Initiating genomic selection in tetraploid potato

Sverrisdóttir, Elsa

Publication date:
2017

Document Version
Publisher's PDF, also known as Version of record

INITIATING GENOMIC SELECTION IN TETRAPLOID POTATO

Elsa Sverrisdóttir
Department of Chemistry and Bioscience
PhD Thesis

AALBORG UNIVERSITY
DENMARK

INITIATING GENOMIC SELECTION IN TETRAPLOID POTATO

by

Elsa Sverrisdóttir

AALBORG UNIVERSITY
DENMARK

Dissertation submitted 20 March 2017

Dissertation submitted: 20-03-2017

PhD supervisor: Professor Kåre Lehmann Nielsen,
Aalborg University, Denmark

PhD committee: Associate Professor Teis Esben Søndergaard (chairman)
Department of Chemistry and Bioscience
Aalborg University, Denmark

Research Scientist Helen Tai
Agriculture and Agri-Food Canada
Fredericton, New Brunswick, Canada

Professor Søren Kjærsgaard Rasmussen
Department of Plant and Environmental Sciences
University of Copenhagen, Denmark

PhD Series: Faculty of Engineering and Science, Aalborg University

ISSN (online): 2446-1636
ISBN (online): 978-87-7112-924-3

Published by:
Aalborg University Press
Skjernvej 4A, 2nd floor
DK – 9220 Aalborg Ø
Phone: +45 99407140
aauf@forlag.aau.dk
forlag.aau.dk

Printed in Denmark by Rosendahls, 2017

Preface

The present thesis was submitted as part of the requirements for attaining the PhD degree at the Faculty of Engineering and Science, Aalborg University.

The thesis is based on work carried out in the period from November 2013 to March 2017. The PhD project is part of GenSAP: Centre for Genomic Selection in Animals and Plants, a collaborative research initiative gathering all relevant Danish breeding companies and research groups as well as world-leading international researchers. The overall aim of GenSAP is to build the foundation for next-generation genomic selection tools for genetic improvement schemes in agricultural plants and animals. GenSAP is partly supported by the Innovation Fund Denmark under grant no. 12-132452.

I have been enrolled at the Department Chemistry and Bioscience, Faculty of Engineering and Science during this project.

The thesis is based on the following papers:

1. **Sverrisdóttir, E**., Byrne, S., Sundmark, E. H. R., Johnsen, H. Ø., Kirk, H. G., Asp, T., Janss, L. and Nielsen, K. L. (2016) *Genomic prediction of starch content and chipping quality in tetraploid potato using genotyping-by-sequencing.* Manuscript submitted to Theoretical and Applied Genetics.
2. **Sverrisdóttir, E**., Sundmark, E. H. R., Johnsen, H. Ø., Kirk, H. G., Asp, T., Janss, L., Bryan, G., and Nielsen, K. L. (2017) *The value of expanding the training population in genomic selection models for tetraploid potato.* Manuscript will be submitted to PLOS ONE.

Furthermore, the following popular science paper is relevant for the project, but not included in the thesis:
Lehmann, K. L. and **Sverrisdóttir**, E. (2016) *Hvorfor DNA-sekventering og kartofler er vigtig - for at brødføde verdens befolkning i fremtiden.* Dansk kemi.

The images used on the cover page were found online and are in the public domain. One exception to this is the image Tractors in Potato Field, which is attributed to user NightTree on Wikipedia.com.

Acknowledgements

First and foremost, I owe my sincerest thanks to my academic advisor, Kåre L. Nielsen, for his invaluable support and guidance, positive outlook on matters, and for inspiring me to pursue an academic career.

I also want to thank Torben Asp and Stephen Byrne who have taken the time to help me with data processing, Luc Janss for introducing me to statistical modelling and guiding me through genomic predictions and basic R knowledge, and Glenn Bryan for his hospitality and for sharing his data with me.

My sincerest thanks go to Hanne Grethe for answering my never-ending questions about phenotyping methods.

I would like to thank my colleagues at the Department of Chemistry and Bioscience and especially the members of the Functional Genomics group. Thanks to Anne, for her help in the lab, and to Mette, for her tremendous support in the beginning of this project and joyful talks in the office – academic or otherwise – and whose presence in the office has been missed since she left. Thanks to my current office buddies, especially to Ea, for fruitful discussions and/or distractions. Additionally, thanks to Mads for computational and bioinformatical guidance as well as endless after-work beers.

Last but not least, I would like to express my special appreciation and thanks to my family and friends for their enormous support, encouragement and patience during this project. A special thanks to Albert for being my everlasting go-to guy for front-page productions. To Martin, for making my life easier and keeping my spirits high, especially when times were difficult. Finally, my most heartfelt thanks to Eva for her everlasting support and assistance during this project, whose help during the last few weeks has been invaluable.

Elsa Sverrisdóttir
March 2017
Aalborg, Denmark

English summary

The world's population is growing rapidly, and by 2050 it is estimated that it will reach its maximum at almost 10 billion people, more than doubling the food demand. Breeding for more space and resource efficient crops is therefore more important than ever, especially since most fertile areas are already intensively cultivated. Potato is the world's most important non-grain food crop and produces approximately twice the amount of calories per hectare compared to cereals. It is one of the most space-efficient food crops and is thus of central importance for global food security.

Potato breeding faces several difficulties, leading to slow breeding gain. The traditional "mate and phenotype" breeding approach is costly and time-consuming, consisting of 10-15 years of hard work, where as much as a million seedlings are screened before a cultivar can successfully be introduced to the market. Molecular breeding has the potential to speed up the breeding process significantly, and the completion of the potato genome sequence has greatly facilitated the application of genomics-assisted breeding technologies. Genomic selection using genome-wide molecular markers is becoming increasingly applicable to crops as the genotyping costs continue to reduce, which makes it an attractive breeding alternative. It allows for prediction of performance of individuals and subsequent selection of breeding candidates in the absence of direct phenotyping.

The overall objective of this thesis was to evaluate the potential of and initiate a genomic selection breeding programme in tetraploid potato. Genotyping-by-sequencing was used to genotype a large number of individuals from three panels. The main panel, the MASPOT population, consisted of 762 individuals derived from a larger population generated from biparental crosses of 18 tetraploid parents in a full diallel crossing design. Additionally, two test panels were established, one panel consisting of elite cultivars and breeding clones from the same breeding station as the MASPOT population, named Test panel DK, and one mapping population from a breeding station in the UK, named Test panel UK. The populations were different in a number of ways. Firstly, the MASPOT population differed from the two test panels in that no selection of individuals had occurred, so that the population encompassed a vast variation of genotypes and phenotypes. Secondly, Test panel UK stood out, being from a different breeding station than the other two, and in particular, being grown and harvested in another country.

Genomic prediction models were generated for starch content, dry matter, chipping quality, yield, and late blight resistance. High cross-validated prediction accuracies were obtained for starch content and dry matter within each population, while prediction accuracies for chipping quality were generally slightly lower. Predictions were significantly lower for yield, due to a low heritability of the trait and a high proportion of non-additive genetic effects. Interestingly, predictions for yield and chipping quality within Test panel UK stood out, being significantly higher than within the two other populations. Two prediction models were constructed for the prediction of late blight resistance: One model in which resistance would be dominated by R gene resistance and one model in which the quantitative resistance was attempted predicted by removing individuals with known R genes. Prediction accuracies were moderate using either model, but predictions were slightly poorer when excluding R genes from the model, signifying the more challenging predictions of quantitative resistance.

Predicting performance of individuals across breeds is of great interest for breeders. In general, low or moderate prediction correlations were obtained across populations. Predictions were particularly poor between populations, in which the traits displayed differences in performance with regard to heritability and within-population prediction accuracies, such as chipping quality and yield predictions between the MASPOT population and Test panel UK. In contrast, predictions were more robust, even across populations, when predicting dry matter, which is a highly heritable trait. However, prediction bias was large in all cases, reflecting a deflation of the scale of predicted values relative to the observed phenotypes.

One of the important parameters in genomic selection is the size of the training population, which together with the cost of phenotyping is the main determinant of cost efficiency. The three populations were thus combined for making an expanded training population. Prediction accuracies were similar to the ones obtained within each model, and thus no gain in prediction accuracy was observed when using the combined population. However, the combined model had maximal prediction accuracy for all populations simultaneously, and could therefore be applied to all populations. This indicates that it is indeed possible to obtain a general potato prediction model if all relevant genotypes are included in the model.

Overall, the results of this study suggest that genomic prediction of important agricultural traits, and hence selection of breeding material by genomic selection, can be obtained with good accuracies within tetraploid potato. Although the most optimal prediction accuracies were obtained when predicting within the same population, the results from combining training populations with genotypes from different populations suggest a promising approach for establishing a broad-application prediction model for the implementation of genomic selection in tetraploid potato breeding programmes.

Danish summary

I år 2050 vil der være næsten 10 milliarder mennesker på jorden, som skal brødfødes, hvilket betyder at der skal produceres dobbelt så mange fødevarer som i dag. Udvikling af nye og bedre afgrøder er derfor vigtigere end nogensinde, og da der reelt ikke er mulighed for at udvide det globale landbrugsarealet, er det især pladseffektivitet der kommer til at være en betydelig parameter for fremtidens landbrug. I den sammenhæng har kartofler en vigtig rolle – kartofler kan nemlig producere ca. dobbelt så mange kalorier per areal sammenlignet med korn, majs og ris, og er dermed en af de mest pladseffektive afgrøder der findes.

En række udfordringer er forbundet med kartoffelforædling, hvilket bl.a. har forårsaget at udbyttet af kartofler ikke er blevet forbedret væsentligt de sidste 50 år; mens udbyttet af korn er forøget med 190% er udbyttet på kartoffelmarker kun steget med 60% i den samme periode. Kartoffelforædling foregår som regel med den klassiske forædlingsmetode, hvor to planter krydses, hvorefter deres afkom bliver selekteret ud fra deres fænotypiske egenskaber. Processen er tidskrævende og omkostningsfuld, da der typisk går 10-15 år fra krydsning til at en færdig sort potentielt kan sættes på markedet. Forædlingsprocessen kan fremskyndes markant ved brug af molekylære teknikker, og kortlægningen af kartoflens i 2011 har givet bedre muligheder for at implementere sådanne metoder i kartoffelforædling. En fremtrædende metode er genomisk selektion, hvor markører der dækker hele genomet bliver brugt til at forudsige plantens egenskaber, uden at fænotypebestemmelse er nødvendig.

Det overordnede formål med denne afhandling var at indlede genomisk selektion i tetraploide kartofler og evaluere potentialet heraf. Genotyper af et stort antal planter blev bestemt med genotyping-by-sequencing. Tre paneler blev etableret. Det største panel, MASPOT populationen, bestod af 762 individer, som blev udvalgt fra en større population genereret fra krydsninger af 18 forældre. Derudover blev to testpaneler etableret. Det ene testpanel, Test panel DK, var en samling af 74 elite- og forædlingssorter fra samme forædlingsstation som MASPOT populationen, hvorimod Test panel UK bestod af 290 sorter fra en forædlingsstation i Storbritannien. Panelerne adskilte sig fra hinanden på en række måder. For det første er ingen selektion foretaget i MASPOT populationen, og den består derfor af et bredt sortiment af genotyper og fænotyper. For det andet kommer Test panel UK ikke blot fra en anden forædlingsstation, men panelet blev også groet og høstet i et andet land.

Genomiske prædiktionsmodeller blev genereret for stivelsesindhold, tørstof, chipping kvalitet, udbytte og skimmelresistens. Prædiktionsværdier var høje for stivelse og tørstof når modellerne blev udført inden for hver population. Værdier for chipping kvalitet var en anelse lavere, mens de var signifikant lavere for prædiktion af udbytte. Dette skyldes bl.a. at udbytte har en lavere arvelighed end de andre træk, og derudover er der mange non-additive genetiske effekter der har betydning for fænotypen. Prædiktioner for udbytte og chipping kvalitet inden for Test panel UK skilte sig ud, idet nøjagtigheden af prædiktioner var betydeligt højere end for de to andre populationer. For skimmelresistens blev der lavet to modeller: En model der hovedsageligt fangede resistens fra dominante R gener, og en model, hvor den kvantitative resistens blev forudsagt, da individer med R gener blev fjernet. Moderate prædiktionsværdier blev opnået i begge tilfælde, men i den sidstnævnte model var værdierne en anelse lavere.

Prædiktioner på tværs af populationer er af stor interesse for forædlere, men generelt blev kun lave eller moderate prædiktionsværdier opnået. Værdier var især lave for prædiktioner mellem populationer hvor trækkene opførte sig væsentlig anderledes, både med hensyn til arvelighed og prædiktioner inden for hver population, som udbytte og chipping kvalitet mellem MASPOT populationen og Test panel UK. Modeller for tørstof var mere robuste, selv på tværs af populationer, men arveligheden for dette træk er høj. Bias af prædiktioner var dog høj i alle tilfælde, som er et udtryk for at skalaen for de forudsagte værdier var smallere i forhold til skalaen for de observerede værdier.

Størrelsen af træningspopulationen der bruges i genomisk selektion har en stor betydning for resultaterne, og sammen med fænotypebestemmelser er den afgørende faktor for omkostningseffektiviteten. De tre populationer blev derfor kombineret for at opnå en stor træningspopulation. Prædiktionsværdier var sammenlignelige med de værdier der blev opnået inden for hver population, og derfor blev prædiktionerne ikke forbedret med den kombinerede model. Dog blev der opnået maksimale prædiktionsværdier alle populationer samtidigt, og dermed kunne den samme model bruges på alle populationer.

Generelt er resultaterne udtryk for at genomisk prædiktion kan udføres med gode resultater for en række vigtige træk i tetraploide kartofler. De bedste resultater blev opnået med modeller inden for hver enkelt population, men resultater opnået med den kombinerede model indikerer at en generel prædiktionsmodel kan udvikles med genotyper fra forskellige populationer til at give en bred anvendelsesmulighed for at initiere genomisk selektion i tetraploide kartofler.

Abbreviations

AUDPC	area under the disease process curve
BLUP	best linear unbiased prediction
BSA	bulk segregant analysis
GBLUP	genomic best linear unbiased prediction
GBS	genotyping-by-sequencing
GEBVs	genomic estimated breeding values
GS	genomic selection
GWAS	genome-wide association study
MAS	Marker Assisted Selection
MASPot	MAShed Potato: Moving potato breeding into the post genome era
NGS	next-generation sequencing
PCA	principal component analysis
QTL	quantitative trait loci
R genes	resistance genes
SNPs	Single nucleotide polymorphisms
SSR	simple sequence repeat

Table of Contents

Aim

The overall aim of this thesis was to initiate a genomic selection breeding programme in tetraploid potato. More specifically, the detailed aims were to:

- Use genotyping-by-sequencing to provide genotypic data of an appropriate number of potato plants
- Construct genomic prediction statistical models for important agricultural traits
- Provide proof of concept for genomic selection in tetraploid potato by predicting performance of un-included potato plants and compare with existing phenotypic data
- Investigate prediction performance across different populations of tetraploid potato

Introduction

The world's population has doubled since the late 1970s to a population size of 7 billion people today. The population continues to increase, and by 2050, it is estimated that it will reach its maximum at around 9.7 billion people (Melorose, Perroy and Careas, 2015), which will more than double the food supply demand (Alexandratos and Bruinsma, 2012; Valin *et al.*, 2013). In addition, an increasing amount of agricultural products are being used for other uses, such as biofuels and bio-based chemicals (Naik *et al.*, 2010), adding even more pressure on agricultural production.

Most fertile areas are already intensively cultivated (Borlaug, 2002), so all extra production will rely on existing farmlands. Therefore, crops with improved space use efficiency will be of utmost importance in order to meet the increasing food demand.

Potato (*Solanum tuberosum* L.) is the third most important food crop worldwide after wheat and rice, with 382 million tons fresh weight of tubers produced in 2014 from 19.1 million hectares of land (FAOSTAT, 2017). Of the most produced food crops, potato is the most space efficient crop, having the potential to produce approximately twice the amount of calories per hectare compared to cereals. In addition, potato plants have the advantage that the storage organs grow underground, and high yielding potato plants can therefore be grown without the need to increase plant strength, contrary to wheat, maize and rice. In a future agricultural scenario, where more food has to be produced from less area, these characteristics make potato an interesting crop.

The potato crop

Potato diversity is vast: approximately 5000 varieties of potatoes are known, most of them belonging to the species *Solanum Tuberosum* (Burlingame, Mouillé and Charrondière, 2009). A general quality of the potato crop is that it is high in carbohydrates, low in fat, and a good source of several antioxidants and vitamins, especially vitamin C and iron. In addition, potatoes are also rich in proteins compared to other roots and tubers (Lutaladio and Castaldi, 2009). However, potatoes only account for around 2% of the world's dietary energy supply (FAOSTAT, 2017). They have a more dominant place in the diets of people in the developed world, where potatoes account for 540 kJ (130 kcal) per person per day, whereas in the developing world, only 170 kJ (41 kcal) of the daily energy supply comes from potatoes (Burlingame, Mouillé and Charrondière, 2009). Up to 60% of the daily potato consumption in developed nations is in the form of processed food, such as potato chips (crisps) and French fries (Kirkman, 2007).

Fundamentally, the need for new cultivars can be divided in two contrasting scenarios. In the developed world, there is a desire for more economically and environmentally sustainable potato production; i.e. there is a need for cultivars providing higher yields at lower costs, as well as pest and disease resistant cultivars in order to reduce the use of pesticides and fungicides. In addition, there is an increased interest for improved nutritional and health benefits, as well as improved flavours (Bradshaw, 2007). In developing countries, however, there is a profound need for an increased and stable potato production to meet food demand. This is perfectly illustrated in the vast variations of potato yield between regions; in Western Europe and the USA the average yields are around 50 tons per hectare, while merely 13 tons are produced per hectare on average in Africa (Barrell *et al.*, 2013). Furthermore, there is a need for cultivars that are tolerable to environmental stresses such as heat, cold, drought and salinity, preferably with improved nutritional and health properties (Bradshaw, 2007).

Important agronomical traits

A number of traits are of general importance for potato breeding, such as tuber yield, starch content, resistance to pests and diseases, and plant maturity (Gebhardt, 2013). In addition, a number of traits are considered for specific production areas, markets, and end uses. For example, fresh market potatoes are phenotyped for traits such as tuber shape, skin and flesh colour, eye depth, texture, taste, cooking type, tissue discolouration after cooking, and susceptibility to bruising.

Most potato quality traits are multigenic and are controlled by many different genes on different chromosomes, for example yield, chip colour and specific gravity (Bonierbale, Plaisted and Tanksley, 1993; Freyre and Douches, 1994; Schäfer-Pregl *et al.*, 1998; Li *et al.*, 2013)

Yield

Tuber yield is an important potato trait for obvious reasons. Yield is probably the most complex polygenic trait and is strongly influenced by environmental conditions and agricultural practices (Schönhals *et al.*, 2016). As previously mentioned,

the average potato yields vary significantly between regions, varying from 13 tons per hectare in developing countries to 50 tons per hectare in developed countries (Barrell *et al.*, 2013). However, the yield in developed countries has not always been as high. For instance, the average potato yields doubled in Great Britain from 22 to 45 tons per hectare in a 44-year period from 1960 to 2003 (Bradshaw, 2006). This increase was the result of a number of contributing factors, such as better disease control, more irrigation and other improvements in agronomy, as well as the development of new cultivars (Bradshaw, 2006). Thus, similar improvements in developing countries can be expected in the future with the appropriate efforts.
Yield has been found to be slightly negatively correlated with tuber starch content (Urbany *et al.*, 2011; Li *et al.*, 2013). In addition, high tuber yield is correlated with late plant maturity (van Eck, 2007; Urbany *et al.*, 2011), most likely since early maturing cultivars will not be able to produce the same amounts of tuber yield in a growing season of 80 days, compared to late maturing cultivars, which are grown for 120 days (van Eck, 2007). Estimations of yield heritability are widely different. Slater, Wilson, *et al.* (2014) estimated the heritability of yield to be 58%, while Urbany *et al.* (2011) estimated the heritability for different populations ranging from 26% to 64%. These differences are a testament to the strong influence environmental factors have on potato tuber yield.

Dry matter and starch content

Freshly harvested potatoes contain about 80% water and 20% dry matter, of which 60-80% is starch (Dale and Mackay, 1994; Lutaladio and Castaldi, 2009). Dry matter content is thus largely determined by starch content, and consequently the two terms are often used as synonyms, although this is not entirely accurate. There is considerable variation both between and within varieties, and dry matter content is subject to considerable influence of the environment, both the growing plant and the stored tubers (Dale and Mackay, 1994).
Tuber starch content and tuber dry matter content are important for both the fresh market and the processing industry. Potato starch is used in a variety of food and non-food products, and in the EU, 18% of the potato production is used for starch production. In Denmark, the number is even higher, as up to 60% is used for starch extraction (Birch *et al.*, 2012). High starch content is obviously desired in this context. Cultivars selected for the production of potato chips (crisps) and French fries also require high dry matter content, however, tubers for table use should only contain moderate dry matter levels in order to prevent disintegration during boiling (Dale and Mackay, 1994).
As for tuber yield, starch content and dry matter has been found to be highly correlated to maturity (van Eck, 2007). However, Johnsen (2015) found no correlation between dry matter content and maturity in a diallel population where no selection had occurred, suggesting that the previously found correlation might be the result of selection bias.
Starch content has been described as a truly quantitative and polygenic trait with genes on every chromosome contributing to phenotypic variation (van Eck, 2007). Slater, Wilson, *et al.* (2014) found the heritability for dry matter content (specific gravity) to be 74%. However, since heritability estimates are a function of genetic and environmental variances, they are appropriate only to the population from which they are derived (Cunningham and Stevenson, 1963).

Chipping quality

Processed potato products, such as potato chips and French fries, constitute up to 60% of the daily potato consumption in the developed world (Kirkman, 2007). One of the most important quality traits for breeding potatoes for potato chips and French fries is the content of reducing sugars. At the high temperatures during frying, reducing sugars undergo the non-enzymatic Maillard reaction with free amino acids, resulting in dark coloured and bitter products as well as the production of carcinogenic acrylamides (Shallenberger, Smith and Treadway, 1959). The concentration of reducing sugars in the tuber is correlated with the colour of a potato chip after frying (Townsend and Hope, 1960; Brown *et al.*, 1990), and the quality of chips and French fries is therefore mainly estimated as the fry colour. Biotic and abiotic stress under cultivation affects the concentration of reducing sugars, as well as the age of the tuber; however, the amount of reducing sugars mainly depends on genotype and storage conditions. In the processing industry, cold storage is necessary to prevent sprouting and fungal attacks, however storage at low temperature (e.g. 4°C) causes reducing sugars to accumulate in the tuber as a result of starch degradation (Isherwood, 1973; Draffehn *et al.*, 2010). This phenomenon, called cold-induced sweetening, is an adaptive response to cold stress.
A large number of genes, markers, and quantitative trait loci (QTL) have been reported to be associated with chipping quality and tuber starch content (Douches and Freyre, 1994; Chen, Salamini and Gebhardt, 2001; Menéndez *et al.*, 2002; D'Hoop *et al.*, 2008; Li *et al.*, 2008, 2013). However, genes strongly associated exclusively with either chipping quality or starch content have not yet been discovered (Li *et al.*, 2013). Invertases have been shown to be important for chipping quality (Draffehn *et al.*, 2010; Baldwin *et al.*, 2011; Schreiber *et al.*, 2014) and more than 20 loci encoding invertases exist in

potato (Schreiber *et al.*, 2014). Vacuolar invertase has been found to be particularly important for this trait (Sowokinos, 2001), and knock-out of this locus leads to high chipping quality (Clasen *et al.*, 2016).

Late blight resistance

Late blight, caused by the oomycete *Phytophthora infestans*, is the major disease of potato. Late blight was the cause of the Irish potato famine in the 1840s, where the potato crop was practically wiped out and millions died or emigrated due to starvation (Bourke, 1993). Late blight spreads rapidly under cool and wet conditions. Resistance to late blight has thus been an important objective in most potato breeding programs during the last century, and tons of fungicides are used every year to control the disease. It has been estimated that around $150 million are spent every year on fungicides for late blight control in Europe (Forbes and Landeo, 2006). However, the use of fungicides is both economically and environmentally unsustainable. Yields of organic potatoes in Europe are typically 50% lower than from conventional fields with losses to late blight causing the most significant reduction, making organic potato farming rather unfeasible (Gianessi and Williams, 2011). In addition, *P. infestans* is genetically highly flexible and new aggressive genotypes develop resistance to the fungicides used for pest control (Chen *et al.*, 2003).

A number of wild potato species, such as *S. demissum*, coevolved with *P. infestans* and have provided the primary germplasm for breeding late blight resistance in cultivated potato (Song *et al.*, 2003). There are two types of late blight resistance. The first is caused by dominant resistance genes (*R* genes) that induce a hypersensitive resistance response upon infection with specific races of *P. infestans*. At least 11 *R* genes from *S. demissum* have been incorporated into various potato cultivars (Umaerus and Umaerus, 1994). The other type is quantitative, race-nonspecific resistance, which is controlled by an unknown number of genes (Gebhardt and Valkonen, 2001). Several *R* genes originating from introgressions of *S. demissum*, *S. bulbocastanum* and *S. berthaultii* have been mapped to potato chromosomes using DNA markers (Leonards-Schippers *et al.*, 1992; El-Kharbotly *et al.*, 1994; Jacobs *et al.*, 1995; Li *et al.*, 1998; Ewing *et al.*, 2000; Naess *et al.*, 2000). *R1* is located on potato chromosome V in a hot spot for resistance to various pathogens (Leonards-Schippers *et al.*, 1992). This region also contains major QTL for resistance to late blight (Leonards-Schippers *et al.*, 1994; Collins *et al.*, 1999; Oberhagemann *et al.*, 1999), which is also the most significant and most reproducible QTL for late blight resistance (Gebhardt and Valkonen, 2001). The genes *R3a*, *R3b*, *R6* and *R7* are located on potato chromosome XI in the same gene segment (El-Kharbotly *et al.*, 1994, 1996, Huang *et al.*, 2004, 2005). *R2* has been mapped to chromosome IV (Li *et al.*, 1998). *Rber,* originating from *S. berthaultii*, and *Rblc*, originating from *S. bulbocastanum*, have been identified and mapped to chromosomes X and VIII, respectively (Ewing *et al.*, 2000; Naess *et al.*, 2000; Song *et al.*, 2003). All of these *R* genes confer race-specific hypersensitive resistance to late blight. Unfortunately, they provide only short-lived resistance in the field as new virulent races of the pathogen rapidly overcome the resistance encoded by single race-specific R genes (Fry and Goodwin, 1997).

Quantitative resistance to late blight is a polygenic trait, and factors controlling quantitative resistance to *P. infestans* have been found on almost every chromosome in potato (Gebhardt and Valkonen, 2001). Furthermore, late maturity is frequently associated with high levels of late blight resistance (Oberhagemann *et al.*, 1999). In fact, two QTL for late blight resistance, on chromosome V and chromosome VI, respectively, are linked with QTL for plant maturity (Collins *et al.*, 1999; Oberhagemann *et al.*, 1999), and resistance to late blight and delayed plant maturity are likely to be pleiotropic effects of the same genes. This is somewhat unfortunate, because ideally a short, high yielding growth season is optimal, since decreasing the growth season will allow more space for double cropping systems.

Traditional potato breeding

Potato breeding is largely conducted through classical selective breeding involving crossing of two heterozygous tetraploid parents followed by years of evaluation and selection, where offspring are propagated vegetatively and phenotyped for quality traits. The parents often have contrasting appealing qualities, and crosses are made with the expectation that a small percentage of the offspring will contain at least some of the desirable traits of both parents. This process allows for the chance to introduce multiple desirable traits, however, it also allows for the incorporation of undesirable traits that must subsequently be selected against (Halterman *et al.*, 2016). The result is 10-15 years of hard work, where as much as a million seedlings are screened before a cultivar can successfully be introduced to the market (Plaisted *et al.*, 1994). Little has changed in the conventional potato breeding process in the past century, and although potato breeding has resulted in the development of highly successful potato cultivars, no significant improvements in the yield potential of potato cultivars have been achieved over the last century (Barrell *et al.*, 2013). In fact, the yield of cereals per area worldwide has increased by 190% in the period between 1961 and 2014, while the yield of potatoes in the same period has only increased by 60% (FAOSTAT, 2017). One of the major difficulties associated with potato breeding is the autotetraploid status (2n = 4x = 48)

of the majority of potato cultivars, resulting in tetrasomic inheritance, thus making the accumulation of desired alleles extremely difficult (Barrell *et al.*, 2013). In addition, potato cultivars are highly heterozygous and suffer from inbreeding depression (Potato Genome Sequencing Consortium *et al.*, 2011). In addition, approximately 40 traits are considered during new cultivar development (Gebhardt, 2013), which adds to the challenges associated with potato breeding, compared to breeding of grains or forages, where less traits need to be taken into account (Slater, Cogan, *et al.*, 2014).
The slow process of traditional breeding has led breeders into seeking alternative breeding methods. Progeny testing is used in many breeding programs to some extent, and in a Scottish breeding program, progeny testing of a number of traits has been successful, although a slow breeding cycle cannot be avoided (Bradshaw, Dale and Mackay, 2003, 2009; Bradshaw, 2007). Best linear unbiased prediction (BLUP), which uses pedigree to estimate breeding values, is extensively practiced in animal breeding. Slater, Wilson, *et al.* (2014) demonstrated that BLUP in potato breeding can result in increased genetic gains for low heritability traits in autotetraploid potato. However, molecular breeding has the potential to speed up the breeding process significantly.

Molecular breeding

A number of molecular techniques have become available for plant breeding with the potential to speed up the breeding process. In particular, the use of diagnostic DNA markers, which associate a genotype with a phenotype, allows for more efficient screening of a large number of plants in a much earlier stage, even before the phenotype can be observed. Breeders can identify genetically elite plants while discarding plants with undesired traits early in the breeding process, and time and money required for conventional plant breeding can be significantly reduced. Such a selection process is called Marker Assisted Selection (MAS). Single nucleotide polymorphisms (SNPs) are the most versatile molecular markers due to their relative abundance across the genome, e.g. in potato, the SNP frequency has been estimated to be approximately 1 per 24 bp (Uitdewilligen *et al.*, 2013). A number of markers have been identified in potato using different identification methods, such as restriction fragment length polymorphism (Gebhardt *et al.*, 1989; Jacobs *et al.*, 1995), amplified fragment length polymorphism (van Eck *et al.*, 1995), simple sequence repeat (SSR) microsatellite (Milbourne *et al.*, 1998; Ghislain *et al.*, 2004; Feingold *et al.*, 2005; Reid *et al.*, 2011), and diversity array technology markers (Śliwka *et al.*, 2012). Most markers available are for disease resistance mediated by R-genes since it is a dominant trait and the presence/absence of the marker is directly indicative of presence/absence of resistance. These markers are therefore ideal for MAS. Tuber quality traits are of great interest, but markers for these traits are often polygenic and complex as described previously. Furthermore, markers for tuber quality traits are often quantitative and thus dependent on allele dosage, which is much more difficult to analyse. Candidate genes associated with chip colour, tuber starch content, and starch yield have been identified (Fischer *et al.*, 2013; Li *et al.*, 2013). Candidate genes that are thought to play a role in the trait of interest are considered to be *perfect markers* (Barrell *et al.*, 2013), as there is no risk of recombination to occur between the marker and the trait. However, identification of these genes requires prior knowledge of the biochemical and physiological processes behind the phenotype. In addition, the gene sequences underpinning these processes and pathways must be known. Since the sequencing of the potato reference genome in 2011 (Potato Genome Sequencing Consortium *et al.*, 2011), the application of molecular marker techniques in potato breeding has become more accessible. However, the autotetraploid nature and high allelic variation of cultivated potato complicates the implementation of markers for potato breeding (Milczarek, Flis and Przetakiewicz, 2011; Barrell *et al.*, 2013; Ramakrishnan *et al.*, 2015).
Only a few cases of implementing MAS in potato breeding have been successful (Rizza *et al.*, 2006; Ottoman *et al.*, 2009; Ortega and Lopez-Vizcon, 2012; Schultz *et al.*, 2012). A major obstacle in molecular plant breeding has been the cost efficiency. Ten years ago, genotyping an entire plant breeding population would have been inconceivable. Fortunately, genotyping costs have reduced significantly over the last ten years, which has made molecular breeding attainable in plant breeding. The main reason for these cost reductions is the introduction of next-generation sequencing (NGS), which has revolutionised genomics research. With the introduction of massively parallel sequencing, the sequencing price per Mb has reduced almost 30,000 times in the last ten years (Wetterstrand, 2017). NGS technologies have allowed for inexpensive genome wide marker discovery, and thus new approaches for sequence-based genotyping have been developed. One promising method is genotyping-by-sequencing (GBS), which is based on reduction of complex genomes with restriction enzymes (Elshire *et al.*, 2011). Coupled with DNA barcoded adapters, multiplex libraries of samples for NGS sequencing can be produced efficiently and inexpensively. It has been implemented in a wide number of species, where it has been proven to be a simple and robust technique, producing tens of thousands to hundreds of thousands of markers (Elshire *et al.*, 2011; Poland *et al.*, 2012). Originally demonstrated in maize and barley (Elshire *et al.*, 2011), GBS has become a powerful

tool in the field of plant breeding. At a low cost, GBS allows plant breeders to genotype entire breeding populations for implementation of genome-wide association studies (GWAS), genomic diversity studies, genetic linkage analysis, and molecular marker discovery for MAS (He *et al.*, 2014). In addition, GBS is perfectly suited for genomic selection (GS) (Poland and Rife, 2012), where genome-wide markers are used to predict performance and select breeding candidates (Meuwissen, Hayes and Goddard, 2001).

Genomic selection

Genomic selection (GS) is a form of MAS that uses genome-wide molecular markers to predict breeding values of individuals (Meuwissen, Hayes and Goddard, 2001). It is assumed that all QTL are in linkage disequilibrium with at least one marker, and that all genetic variance can be explained by the markers (Meuwissen, Hayes and Goddard, 2001; Goddard and Hayes, 2007). Marker effects are estimated from phenotypes and genotypes of a training population, after which prediction models can be constructed, based on the association between genotypes and phenotypes, see Figure 1. Subsequently, genomic estimated breeding values (GEBVs) can be estimated for a breeding population using only genotypic data. These GEBVs can then be used to select good breeding candidates (Meuwissen, Hayes and Goddard, 2001; Goddard and Hayes, 2007; Heffner, Sorrells and Jannink, 2009).

GS differs from MAS in that it jointly analyses all marker data and thus captures all the genetic variance, whereas MAS relies on a limited number of QTL. MAS is therefore best suited for traits with a few major-effect genes, and not for traits where the genetic variation is the result of a large number of loci of small effect (Dekkers and Hospital, 2002), e.g. yield and other qualitative traits (Heffner, Sorrells and Jannink, 2009). GS was first implemented in breeding programs for dairy cattle (B. Hayes *et al.*, 2009; Luan *et al.*, 2009; VanRaden *et al.*, 2009; Wiggans, VanRaden and Cooper, 2011; Boichard *et al.*, 2012), and a number of studies report promising results for GS in both animal and plant species, such as pigs (Wellmann *et al.*, 2013), sheep (Daetwyler *et al.*, 2010), chicken (Wolc *et al.*, 2011, 2015), wheat and maize (Crossa *et al.*, 2010; Riedelsheimer *et al.*, 2012; Arruda *et al.*, 2015), eucalyptus (Grattapaglia *et al.*, 2011), pine (Resende *et al.*, 2012), and soybean (Jarquín *et al.*, 2014; Ma *et al.*, 2016).

Recently, GS simulation studies have been made in tetraploid potato (Slater *et al.*, 2016), demonstrating that GS is a promising strategy for potato breeding. Using a training population of 5000 individuals, Slater *et al.* (2016) estimated a genetic gain over 20 years almost nine times as high as the expected genetic gain using phenotypic selection, and even when using merely 500 individuals, genetic gain is expected to more than double compared to phenotypic selection. These estimations were made for breeders' visual preference, a trait with low heritability; however, even higher genetic gains were estimated for other traits with higher heritability, such as maturity and boiling colour (Slater *et al.*, 2016). Trait heritability is indeed one of the parameters listed as having an impact on prediction accuracy in GS (B. J. Hayes *et al.*, 2009; Goddard, 2009). In addition, the distribution of QTL effects is important for the success of GS. Heritability and QTL distribution cannot be modified. However, other parameters that significantly influence GS accuracy are under the control of the experimenters or breeders, namely the level of linkage disequilibrium between the markers and the QTL, which can be affected by increasing marker density, and the number of individuals in the training population (Meuwissen, Hayes and Goddard, 2001; Calus *et al.*, 2008). Prediction accuracy can be increased by increasing the size of the training population, as more individuals will result in more observations per SNP allele and thus the SNP effects will be estimated more accurately. Additionally, for traits with low heritability, large numbers of individuals are required in order to achieve high prediction accuracies, while fewer individuals are necessary for high-heritability traits (B. J. Hayes *et al.*, 2009). Similarly, when dealing with traits where a large number of QTL of very small effects contribute to variation, breeders must compensate for loss in prediction by increasing the number of phenotypic records. In addition, to maximize GEBV accuracy, the training population must be representative of selection candidates in the breeding program to which GS will be applied (Heffner, Sorrells and Jannink, 2009).

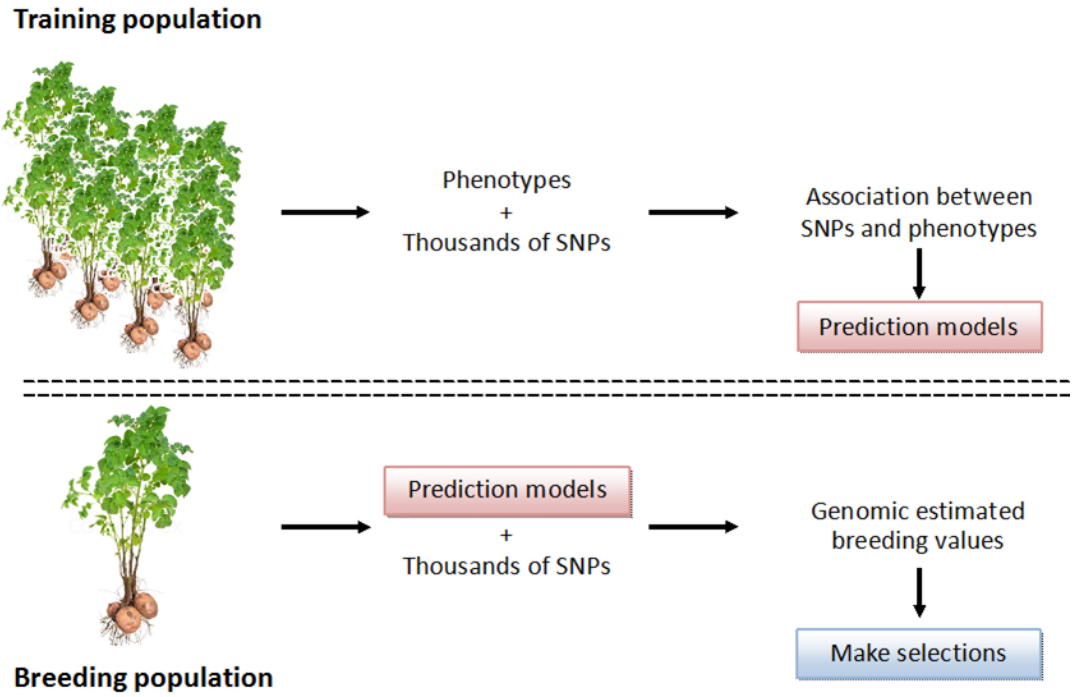

Figure 1 Depiction of the genomic selection process. A training population is genotyped and phenotyped, after which prediction models can be constructed by establishing associations between SNPs and phenotypes. Genomic estimated breeding values can subsequently be calculated for a breeding population based on SNPs, after which selections can be made.

Compared to a traditional breeding program with high selection intensity, Slater *et al.* (2016) estimated that when using a GS training population of 2000 individuals, cost savings around 570,000 Australian dollars could be made, which includes savings due to the smaller breeding population, smaller phenotyping trials, and reduced need for repeat trials to confirm results. This results in breakeven genotyping costs at 100 Australian dollars per genotype in order for GS to be cost-effective (Slater *et al.*, 2016). However, genotyping can easily be made much less expensive, which results in total revenue for GS potato breeding programmes. Furthermore, with the potential genetic gains accompanied with GS, it appears to be a promising approach for potato breeding.

Presentation of main methods

Phenotypes and genotypes

Three different populations were used in this study. The MASPOT population is originally a mapping population established at the breeding station in Vandel, Denmark, and consists of roughly 5000 offspring that were generated by systematic cross-pollination of 18 distinct potato cultivars, which were either established varieties or advanced breeding clones. The population was developed for a project between Aalborg University, LKF-Vandel, and CLC bio, Denmark, running from 2012-2017 and called MAShed Potato: Moving potato breeding into the post genome era (MASPot). For this project, 762 clones were randomly chosen from the original 5000 offspring, and in this thesis, this subset of the original population is called the MASPOT population. The offspring were planted around mid-April to mid-May and harvested in late August in field trials at Vandel, Denmark in 2013 and again in 2014 in duplicates. Test panel DK consists of 92 individuals, including the 18 parents to the MASPOT population, that were selected from a mixture of elite cultivars and breeding clones. However, in Paper 1, the parents were not included in the test panel, which therefore only consisted of 74 individuals. The cultivars were grown, harvested and phenotyped for a number of years in Vandel, Denmark, and phenotype assessments were done as described for the MASPOT population. Test panel UK, developed for Paper 2, consisted of 292 breeding clones and cultivars grown, harvested, and phenotyped at two different sites in the UK (Cambridge and York) in 2012 and 2013 in two replicates. A description of the phenotyping experiments of chipping quality and starch content or dry matter content for the three populations can be found in the methods section in Paper 1 and Paper 2.

Besides the results from the included papers, results from genomic predictions on yield and late blight resistance will be presented in this thesis. For the MASPOT population and Test panel DK, yield was measured in the field at harvest as total weight of five tubers from each clone in two replicates, i.e. 2x5 tubers each. As the plots may vary from year to year and depending on the application, the weight is converted into hkg/ha values, assuming approximately 40,000 plants per ha. Yield for Test panel UK was measured as mean weight (kg) per plot, having 14 plants in each plot. In order to correlate the data to the data for the Danish panels, several methods were considered, the simplest one being to convert the weight/plot from the UK data to hkg/ha values. However, this method assumes that the same number of plants is planted per ha and this information is not available. Therefore, a correlation was made between five individuals present in both Test panel DK and Test panel UK, and the correlation was then used to convert the UK data to hkg/ha.

Late blight resistance was assessed as foliage resistance in field assessments. Tubers were planted in mid-April to mid-May next to an infector row planted with susceptible cultivars that were infected around 10 July. The *P. infestans* isolate used for infection was collected in the fields from plants infected by local and naturally occurring epidemics in the preceding year, and the race structure was thus representative of the local conditions. Degree of infection was assessed twice a week as the area of attacked leaf tissue until a reference cultivar (Robijn) was 50% attacked. The area under the disease process curve (AUDPC) was then calculated for each cultivar:

$$AUDPC_k = \sum_{i=1}^{n} A_{ki} t_i$$

where A_{ki} is the area of infected leaf tissue (0-100) for cultivar k at assessment i, and t_i is the number of days since last assessment. An AUDPC value relative to the mean AUDPC value of the entire year (μ_{AUDPC}) was calculated for each cultivar:

$$AUDPC_{relative,k} = \frac{AUDPC_k}{\mu_{AUDPC}} \cdot 100$$

Resistance was then assessed on a scale from 1 to 9 ($AUDPC$ >210 → 1, $AUDPC$ ≤210 → 2, $AUDPC$ ≤180 → 3, $AUDPC$ ≤150 → 4, $AUDPC$ ≤120 → 5, $AUDPC$ ≤90 → 6, $AUDPC$ ≤70 → 7, $AUDPC$ ≤45 → 8, $AUDPC$ ≤20 → 9).

Late blight resistance in potato is both found as qualitative resistance controlled by dominant R genes and quantitative resistance controlled by an unknown number of genes. In order to separate those resistance types, two separate prediction models were constructed for late blight resistance; a prediction model where individuals known to have R genes for late

blight or offspring that have a parent with R genes are removed in order to capture the quantitative resistance, and a model where no selection of individuals has occurred. In the MASPOT population, three of the parents are known to have R genes, namely 05-GQE-02, Kuras, and Sarpo Mira, however, it is not impossible that there are more individuals containing partially broken R genes in the MASPOT population or in the Test panel.

A detailed description of the DNA extraction, GBS protocol and adapter design can be found in the method section in Paper 1 and Paper 2. A list of adapter sequences is given in Online Resource 3 belonging to Paper 1.

Statistical models

The idea behind GS is simple but powerful: regress phenotypes on all available markers without any marker selection using a linear model (Meuwissen, Hayes and Goddard, 2001). However, with high-density SNP panels the number of markers (*p*) can easily exceed the number of records (*n*) by multiple factors, and fitting this *large-p, small-n* regression requires the use of some type of variable selection or shrinkage estimation procedure (de los Campos *et al.*, 2013). Several shrinkage estimation methods have been proposed, however, the choice of model is important for the success of genomic prediction and the most appropriate model should be chosen for each purpose, i.e. the model that results in highest prediction accuracy with consideration of model complexity and computation requirements (Heffner, Sorrells and Jannink, 2009). For example, linear methods assume that all markers contribute equally to genetic variation, i.e. there are no major genes, while nonlinear (Bayesian) prediction methods assume that the prior distribution of marker or QTL effects is not equal, and that major genes exists giving rise to unequally distributed genetic variance across the chromosomes (VanRaden, 2008). Thus, predictions of traits that are affected by major genes may be better when using nonlinear prediction methods, while either method is sufficient for traits affected by many small-effect QTL.

In this thesis, genomic best linear unbiased prediction (GBLUP) was primarily used for predictions. GBLUP estimates GEBVs directly by using a genomic relationship matrix (VanRaden, 2008). The prior distribution of GBLUP algorithm assumes an equal variance across each locus, however this is not an accurate assumption when the number of QTL is small (Meuwissen and Goddard, 2010). It is widely assumed that GBLUP performs a homogeneous shrinkage of marker effects; however it has been found that marker shrinkage is dependent on allele frequency and sample size, while it is independent on effect-size (Gianola, 2013). The basic model can be described as follows:

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{g} + \boldsymbol{e}$$

where $\boldsymbol{y}$ is a vector of phenotypes, $\boldsymbol{\mu}$ is the mean, $\boldsymbol{e}$ is a vector of random normal deviates, and $\boldsymbol{g}$ is a vector of random genomic breeding values with the distribution:

$$\boldsymbol{g} \sim N\left(0, \boldsymbol{G}\sigma_g^2\right)$$

where $\boldsymbol{G}$ is the genomic relationship matrix and σ_g^2 is the genetic variance for the model. The genomic relationship matrix was created from the genotype matrix ($\mathbf{Z}$) according to VanRaden (VanRaden, 2008) with $\mathbf{Z}$ containing allele frequencies for each sample and SNP computed from sequence data (Ashraf *et al.*, 2016). The allele frequencies were values between 0 and 1 calculated as the ratio between allele counts of the alternative allele and the total allele count:

$$AF = \frac{AC_{alt}}{AC_{ref} + AC_{alt}}$$

The allele frequencies were corrected for missing data using the following correction as described by VanRaden (2008, p. 4420):

$$w_i = \sqrt{\frac{\sum p_k(1-p_k)\ over\ all\ loci}{\sum p_k(1-p_k)\ over\ only\ non-missing\ loci}}$$

where p_k is the mean allele frequency at locus k. The genotype matrix was centred and adjusted for missing values as described by Ashraf *et al.* (2016), after which missing values were set to zero, corresponding to a mean imputation for missing data.

$$\boldsymbol{Z}_{ik} = (\boldsymbol{X}_{ik} - p_k) \cdot w_i$$

The genomic relationship matrix was scaled using global scaling (VanRaden method 1) (VanRaden, 2008).

$$\mathbf{G} = \frac{\mathbf{Z}'\mathbf{Z}}{0.25 \sum p_k(1 - p_k)}$$

where $0.25 \sum p_k(1 - p_k)$ is the sum of genotype variance and also the average diagonal of $\mathbf{Z}'\mathbf{Z}$.

Predictions were also done with two Bayesian models, BayesA and BayesC. Bayesian models allow the markers to explain different degrees of variation. In BayesA, each marker effect is drawn from a normal distribution with its own variance, allowing the marker to be shrunken toward zero to a different degree (Meuwissen, Hayes and Goddard, 2001).

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{X}b + \boldsymbol{e}$$

where every marker effect is assumed to have its own variance parameter:

$$b_i \sim N(0, \sigma_{bi}^2)$$

and where the prior distribution for all variances is a scaled inverted chi-square distribution:

$$\sigma_{bi}^2 \sim \chi^{-2}(\upsilon, S)$$

where υ is the number of degrees of freedom and S is a scale parameter.
BayesC assumes the marker effects to be a mixture, with most marker effects to be zero, and a (usually) smaller part of markers to be nonzero. There is a common marker effect variance for all markers with nonzero effect.

$$b = \begin{cases} 0 \\ \sim N(0, \sigma_b^2) \end{cases}$$

All models were performed with the BGLR package in R 3.3.3 (de los Campos and Perez Rodriguez, 2015; R Core Team, 2015) with default settings for priors. Twelve thousand iterations were used and a burn-in setting of 2000. In almost all cases (except when predicting across populations with different training and validation populations), k-fold random cross-validation schemes were applied. The number of folds was chosen appropriately according to the number of individuals in the population, i.e. 8-fold cross-validation was used within the MASPOT population of 755 individuals. Briefly, the data were randomly divided into eight groups and one group was then used as validation set while the remaining seven groups were used as training population. The process was repeated, each time with another group as validation set, until predictions had been obtained for all individuals. Each analysis was repeated with 10 different cross-validation groupings and the average GEBV over the 10 samplings was taken. The accuracy of the GEBVs was determined as the Pearson correlation coefficient between the GEBVs and the observed phenotypes, described in this thesis as prediction correlation:

$$r(GEBV : y)$$

In one analysis, a leave-sibs-out cross-validation system was applied, in which the MASPOT population was split into groups of full- and half-sibs. Essentially, the 18 parents used for the MASPOT population were split into nine pairs, and the offspring were then divided into nine groups based on the parents, such that each group contained all offspring to one or both parents of the pair in question. Predictions were performed for every group, while making sure that full- and half-sibs were not present in both the training population and the validation population simultaneously. Most individuals were

present in two groups, and thus present in the validation population twice, and in which case, the average GEBV was calculated for further analysis.

Bulk segregant analysis

Molecular breeding has been focused around identifying markers for MAS. In the Functional Genomics group at Aalborg University, marker selection has revolved around bulk segregant analysis (BSA) using the same populations and phenotype data. While GS is a genome-wide marker approach, the rationale of BSA is to identify QTL regions and select markers with the highest predictive power for relevant traits. It is presently unknown which approach is the more powerful, and it is therefore of interest to compare the BSA approach with GS, and in particular, with the GBS approach. A *simulated* BSA study was thus performed on the GBS data and compared with results from a regular BSA study on dry matter. Both studies were performed on the MASPOT population; the regular BSA study selected bulks from the original MASPOT population of ~5,000 individuals, while the bulks for the simulated study with GBS data were selected from the reduced population of 755 individuals (otherwise referred to as the MASPOT population in this thesis). For each case, two bulks of 96 individuals each were gathered by selecting the lowest and the highest performing individuals with regard to dry matter content.
The bulks selected from the GBS data were compared by performing an independent two-sample t-test on average allele frequencies with a null hypothesis that the population means are equal: $H_0: \mu_1 = \mu_2$ and $H_1: \mu_1 \neq \mu_2$, and assuming that each of the two bulks follow a normal distribution with the same variance. The t-test was performed in R with the t.test function (R Core Team, 2015). p-values were then combined with Fisher's method using the sumlog function of the metap package in R (Dewey, 2017). Fisher's method combines p-values into one test statistic (χ^2) using the formula:

$$\chi^2_{2k} \sim -2\sum_{i=1}^{k} ln(p_i)$$

where p_i is the p-value for the i^{th} hypothesis test and k is the number of tests being combined. Markers were binned in 1000 markers each, and significance levels were estimated for each chromosome as the false discovery rate (FDR) as described by Magwene, Willis and Kelly (2011).

Bulks selected for BSA on the original MASPOT population were sequenced and χ^2-tests were used to compare marker distributions between the bulks (best bulk against poor bulk, poor bulk against best bulk). Markers were binned in 12,500 markers each with a step size of 500, and significance thresholds were determined for each chromosome with FDR. More detailed information on the BSA can be found in (Johnsen, 2015) (available on request).

Genome-wide association study

A GWAS was performed on the GBS data by fitting the phenotypic data to each SNP with the genomic relationship matrix as covariance. The basic GBLUP model described before was used and marker effects (β) were added one at a time:

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{x}\beta + \boldsymbol{g} + \boldsymbol{e}$$

where $\boldsymbol{x}$ is an $n \times 1$ marker genotype vector for n individuals at a marker locus and β is the marker effect. GWAS was performed with the regress package in R (Clifford and McCullagh, 2006, 2014). Significance thresholds for p-values were determined for each chromosome with FDR.

Summary of results from papers

This section will consist of a summary of the results from the two papers included in this thesis:

1. Genomic prediction of starch content and chipping quality in tetraploid potato using genotyping-by-sequencing (Paper 1)
2. The value of expanding the training population in genomic selection models for tetraploid potato (Paper 2)

Genomic prediction of starch content and chipping quality in tetraploid potato using genotyping-by-sequencing

In this study, we described the results of genomic prediction of tetraploid potato for starch content and chipping quality. We applied GBS to a training population of 762 individuals, called the MASPOT population, and estimated GEBVs with 171,859 markers using three statistical models: GBLUP, BayesA, and BayesC. 8-fold random cross-validation was used to estimate GEBVs within the MASPOT population. In addition, GEBVs were estimated for a test panel of 74 breeding clones for model validation, using the MASPOT population as training population. Finally, the two populations were combined and GEBVs were estimated with 8-fold random cross-validation.

The composition of the MASPOT population and the test panel was quite different. The MASPOT population was not subjected to any selection, while the test panel consisted of a selection of breeding clones and elite cultivars that have been selected for generations. I addition, the MASPOT population was created from a full diallel crossing design with 18 parents, that were selected to be as unrelated as possible in the inbred elite potato germplasm in order to create as diverse offspring as possible. This difference in selection process was clearly reflected in the phenotypic distributions of the two populations, especially for starch content, where the main part of the test panel had very high starch content, while the mean starch content for the MASPOT population was moderate (Fig. 1 in Paper 1).

The high genetic diversity in the MASPOT population was also illustrated with principal component analysis (PCA), in which the individuals were scattered around the plot, although the explained variance of each principal component was low (Figure 2). The test panel, which was expected to have a lower genetic diversity due to the continued selection of elite cultivars, was grouped together in the PCA plot. The test panel did not seem to be as genetically different from the MASPOT population as was expected, since the two populations overlapped in the PCA plot, however, the diversity within the two populations was clearly different.

No clear population structure was seen in the PCA plot nor in the heat map of the genomic relationship matrix (Figure 3). Nonetheless, there was some family structure as expected due to the design of the MASPOT population. Family structure could be detected in the heat map as clusters around the diagonal, indicating the stronger genetic relationships between full and half siblings. Population structure, or admixture, on the other hand was distinct from family structure. Population structure is found in populations, where all parents are not mated randomly, e.g. a population with 'sub-lines' in which parents are mated (mostly) within sub-line. These would be represented as off-diagonal clusters in Figure 3 and this was not observed. Population structure complicates genomic prediction for a number of reasons, in particular, that such populations do not conform well to the Hardy Weinberg law, which is a fundamental assumption of genomic prediction models.

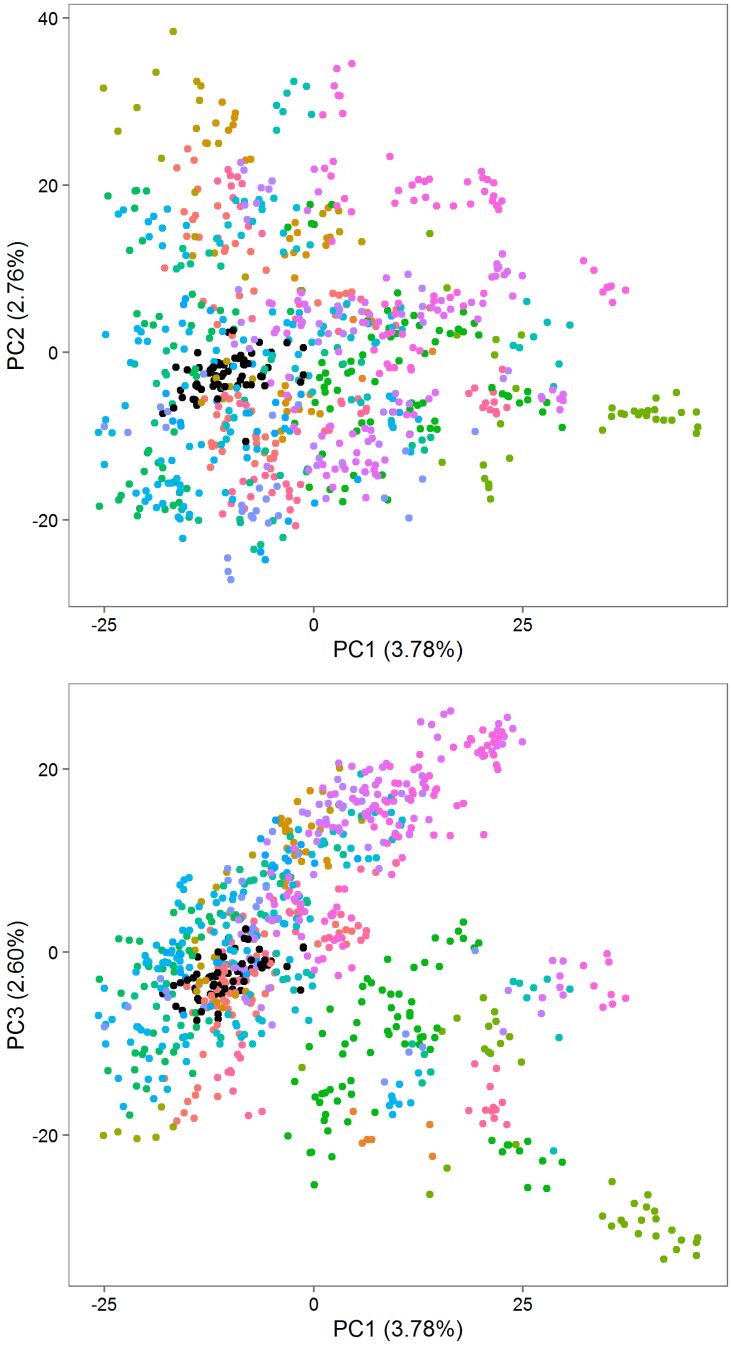

Figure 2 Principal component analysis (PCA) of allele frequency data. Top: Principal component 1 and 2. Bottom: Principal component 1 and 3. Black indicates individuals in the test panel, while the rest of the colours correspond to the mother of the offspring in the MASPOT population.

The test panel seemed to have a lower genetic relationship with the MASPOT population; but there were clearly individuals that shared some alleles with individuals in the MASPOT population. Given that the test panel consisted of individuals that have been selected for specific traits for generations, it is possible that the gene pool of the elite cultivars had become somewhat restricted, resulting from indirect selection of the same alleles. Alternatively, given that nearly all the clones in the test panel clustered together, they may simply have been less related to any of the 18 MASPOT parents than to each other. Narrow-sense heritability was estimated for the MASPOT population with two methods: From genomic and phenotypic variances and from pedigree data, i.e. parent-offspring regressions. Heritabilities estimated for chipping quality ranged from 65% for the genomic heritability to 74% for the pedigree-estimated heritability. For starch content, however, the estimated heritability differed markedly by the two methods; 40% estimated from genomic data, and 97% estimated from parent-

offspring regression. Starch content was considered to be a highly quantitative trait (van Eck, 2007), and was therefore unlikely as highly heritable as estimated from pedigree data. Still, heritability is a function of genetic and environmental variances, and it is therefore appropriate only to the population from which it is derived (Cunningham and Stevenson, 1963). In the MASPOT population, the genetic variance was quite large due to the diallel crossing design and the fact that no selection had occurred in the population, and as long as the environmental variance remained relatively constant, the heritability increased. Slater *et al.* (2014) estimated the heritability for specific gravity to be 74%. Therefore, the heritability of 40% estimated from genomic data seems to be underestimated. The genomic heritability is defined as the proportion of variance of phenotypes explained by the regression on available markers, and given the high diversity of potato with one SNP per 24 bp (Uitdewilligen *et al.*, 2013), the genomic diversity was likely to be underestimated using only 171,859 markers out of a genome size of 8.4 Mbp. In other words, the markers were not in unique 1:1 linkage with a single allele. From this and given that the phenotypic variance was quite high for the MASPOT population, it follows that the genomic heritability would be underestimated by the genomic method.

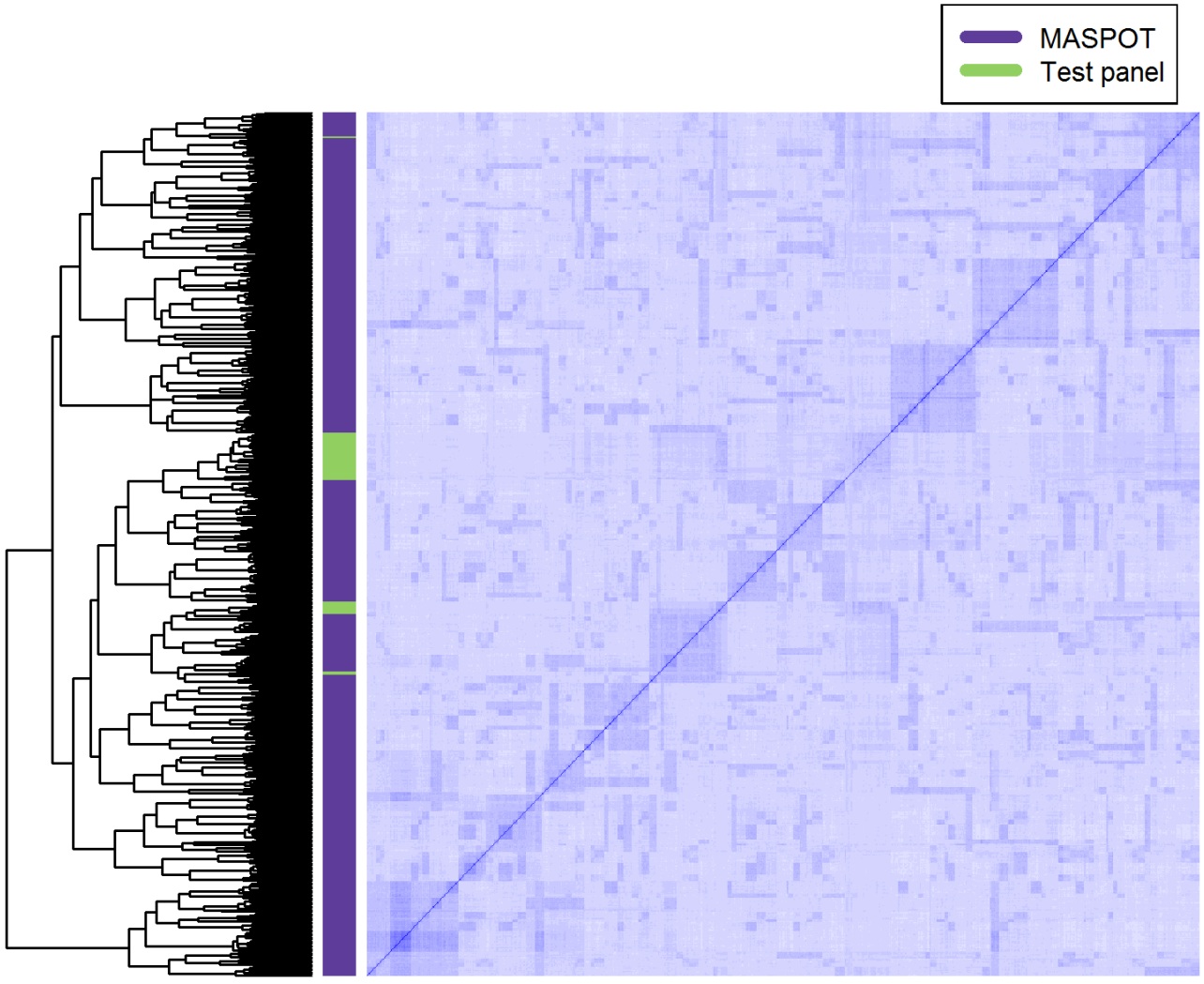

Figure 3 Heat map of the genomic relationship matrix for the 755 offspring in the MASPOT population (purple marking in the left panel) and the 63 individuals in the test panel (green marking). The matrix is obtained from 171,859 markers. Rows and columns represent each individual. The absence of obvious high intensity off-diagonal clusters indicates the absence of population structure.

High cross-validated prediction correlations were obtained when predicting starch content within the MASPOT population, while prediction correlations for chipping quality were moderate (Table 1). The predictions were unbiased, estimated from the slope of the regression line between the observed (y) and the predicted values (x), where β of 1 indicates no bias. As expected, prediction accuracies were lower when predicting out-of-population, i.e. when using the MASPOT population as a

training population and estimating GEBVs for the test panel. The largest difference was seen for starch content predictions, where accuracies dropped from 0.73-0.74 within the MASPOT population to 0.39-0.40 for the test panel, reflecting profound differences in the genetics underpinning starch content between the two populations. When combining the populations and predicting starch content within the combined set with 8-fold random cross-validation, prediction correlations for the test panel were improved to 0.62-0.63, while predictions for the MASPOT population were identical to those obtained when using only the MASPOT population as training population. This suggests that a model can be used to make good predictions in two genetically and phenotypically different populations, as long as representatives from each population are present in the training population.

Similar to starch content, prediction correlations were also lower when predicting chipping quality in the test panel using the MASPOT population as training population. However, the most notable difference was seen in the bias (1.99-2.15), which was much larger than it was within the MASPOT population, meaning that the highest GEBVs were underestimated relative to the observed phenotypic values, while the lowest GEBVs were overestimated. The bias was improved when using the combined populations as training population, although prediction correlations remained the same as when using the MASPOT population as training population only. Predictions with the combined set were made with 8-fold cross-validation, where the individuals were divided in eight equally sized groups selected randomly, and this selection was repeated 10 times. For each cross-validation set, prediction correlations varied considerably when predicting chipping quality (0.29-0.58), while correlations for starch content were more robust (0.59-0.65). This is most likely a reflection of the varying number of phenotypes available for each trait. While starch content data was available for all individuals that remained after filtering, i.e. 755 individuals in the MASPOT population and 63 individuals in the test panel, chipping quality data was available for only 524 individuals belonging to the MASPOT population, and for less than half of the test panel, or merely 30 individuals. Since the cross-validation sets were completely random, chipping quality predictions in the test panel were obviously much more sensitive to the groupings, having only 30 phenotypes among 554 to be divided in eight groups.

The three models tested in the study – GBLUP, BayesA and BayesC, performed very similarly for both traits. Bayesian methods are believed to have an advantage over GBLUP when there are QTL with moderate to large effects on the trait. Starch content and chipping quality are both considered highly polygenic traits, and it is thus not surprising that the models performed similarly.

Table 1 Mean prediction correlations and bias found with BayesA, BayesC, and GBLUP over 10 repeats. Predictions were done using only the MASPOT population or both the MASPOT population and the test panel (combined) to train the model. Predictions made with the combined model as well as predictions made within the MASPOT population were done using 8-fold cross-validation (*). The number of phenotypes available in each case is indicated with square brackets.

Trait / Test set	Training set	BayesA		BayesC		GBLUP	
		Correlation	Slope	Correlation	Slope	Correlation	Slope
Chipping quality							
MASPOT [524]	MASPOT*	0.55	1.03	0.55	1.06	0.55	1.06
Test panel [30]	MASPOT	0.41	1.99	0.43	2.15	0.42	2.13
Combined [554]	Combined*	0.55	1.04	0.55	1.07	0.55	1.07
MASPOT [524]	Combined*	0.55	1.04	0.55	1.06	0.55	1.07
Test panel [30]	Combined*	0.44	1.36	0.44	1.47	0.44	1.50
Starch content							
MASPOT [755]	MASPOT*	0.73	1.01	0.74	1.03	0.74	1.03
Test panel [63]	MASPOT	0.40	1.13	0.40	1.18	0.39	1.16
Combined [818]	Combined*	0.82	1.05	0.82	1.05	0.82	1.05
MASPOT [755]	Combined*	0.74	0.98	0.74	0.99	0.74	0.99
Test panel [63]	Combined*	0.63	0.97	0.62	0.97	0.62	0.97

The value of expanding the training population in genomic selection models for tetraploid potato

In this study, we expanded the population from Paper 1 in order to study genomic prediction across breeds. In addition to the MASPOT population and the test panel studied in Paper 1, we also genotyped a test panel that was grown and harvested in the UK. Furthermore, we added the 18 parents that were used to generate the MASPOT population to the test panel from Paper 1, called Test panel DK in this study. Genomic prediction models were generated with GBLUP using 167,637 markers obtained with GBS. GEBVs were calculated within each of the three populations as well as across the populations. Also, the three populations were combined to make a large mixed training population, which was used to calculate GEBVs for each individual.

In this study, we identified certain disadvantages connected with GBS. The populations were analysed with PCA, and at first glance, Test panel UK seemed to be highly distinct from the two Danish populations. Further investigation made it evident that the differences were caused by a high number of missing data between the data sets. The PCA plots looked vastly different after more stringent filtering, where only SNPs with <1% missing data were retained, which caused the test panels to overlap the MASPOT population (Fig. 2 in Paper 2). Test panel DK and Test panel UK still showed distinct groupings, while the MASPOT population was somewhat scattered, showing larger genetic diversity within the population.

High cross-validated prediction correlations were obtained for dry matter predictions within each population (Table 2). The same was also the case for chipping quality predictions except for within Test panel DK, although, similarly to Paper 1, only a few chipping quality phenotypes were available for this panel, which is likely the cause for the poor predictions. GEBVs estimated within Test panel DK and Test panel UK, respectively, were slightly biased (1.41-1.59) for both traits, while only insignificant biases were observed within the MASPOT population. This can be caused by a number of things: For one, the MASPOT population was significantly larger than both Test panel UK and Test panel DK, potentially giving rise to more robust prediction models with respect to both bias and GEBV accuracy. In addition, the phenotypic range in the MASPOT population was larger than for both Test panel UK and Test panel DK, especially for dry matter content. Particularly the poor performing individuals are not present in the test panels, and this may lead to a steeper slope, giving larger biases. Prediction correlations obtained with the combined model were equal to the ones obtained within each population, though again with chipping quality in Test panel DK as an exception. This supports that good prediction accuracies can be obtained in widely different populations using only one prediction model, as long as representatives from each population are present in the training population, further confirming the results obtained in Paper 1. Although predictions did not improve by expanding the training population, the results certainly suggest that a general model with a broader applicability can be generated and used across breeds. However, the prediction ability across populations, using one population as training set and another as test population, was generally lower, and the observed bias was larger.

Table 2 Mean prediction correlations and bias found with GBLUP over 10 repeats with 167,637 markers, using the three populations separately and combined for modelling. The population used for training the model is listed horizontally while the predicted population is listed vertically. Bias is listed in brackets. For each within-population test (MASPOT model to predict MASPOT etc.) and the combined model, k-fold cross-validation schemes were used. The number of phenotypes available in each case is indicated with square brackets. Bold lettering indicates within-population predictions, where the same population was used for training and test population.

Prediction set / Training set	MASPOT	Test panel DK	Test panel UK	Combined
Chipping quality				
MASPOT [524]	**0.56 [1.09]**	0.35 [1.23]	0.32 [0.54]	0.57 [0.96]
Test panel DK [40]	0.60 [2.52]	**0.34 [1.41]**	0.67 [2.06]	0.57 [1.42]
Test panel UK [290]	0.43 [2.08]	0.37 [3.89]	**0.79 [1.59]**	0.79 [1.39]
Dry matter				
MASPOT [755]	**0.75 [1.04]**	0.65 [1.47]	0.62 [1.55]	0.76 [0.98]
Test panel DK [80]	0.68 [1.83]	**0.81 [1.48]**	0.62 [2.98]	0.83 [1.06]
Test panel UK [290]	0.58 [1.67]	0.36 [2.07]	**0.71 [1.57]**	0.75 [1.28]

In order to investigate the influence of the large number of missing data points on the GEBV estimations, predictions were also made with the stringently filtered marker set consisting of 7,800 markers. As mentioned, these markers were selected based on <1% missing data, and only individuals containing <10% missing data were retained. For comparison, predictions were also made with 7,800 randomly selected markers. Surprisingly, only insignificant differences were found between prediction correlations for the larger and the small marker sets (see Table 2 and 3 in Paper 2). For predictions with the "cherry-picked" marker set, i.e. the marker set with small amount of missing data, the biases decreased in most cases, while predictions performed with the randomly selected markers were quite similar to the ones obtained with the larger set. This suggests that only a small part of the 167,637 markers were used for prediction, making the rest of the markers redundant. Indeed, as discussed on page 36 (Effect of marker number on prediction accuracy), approximately 10,000 markers seem to be sufficient to make good predictions for dry matter and chipping quality. Too many markers tend to produce overfitted models, and to avoid that, it has been suggested that when using increased marker density, the training population size must be scaled with marker numbers in order to capture the additional information provided by increased marker density. Otherwise, any positive effect on accuracy from increasing marker numbers can be constrained by the marker population size (Muir, 2007; Meuwissen, 2009; Lorenz *et al.*, 2011).

Additional results

In this chapter, additional results that have been generated during this PhD will be presented. These results will be published in the future, and a list and descriptions of planned publications can be found at the end of this thesis.

Genomic predictions of yield and late blight resistance

Genomic prediction models were generated with GBLUP for prediction of yield and late blight resistance. Predictions of yield were performed using the same setup as described in Paper 2, using the MASPOT population of 755 individuals, Test panel DK of 80 individuals, including the 18 parents that were used to generate the MASPOT population, and finally Test panel UK of 292 individuals. Late blight resistance data was not available for Test panel UK, and thus late blight resistance predictions were only performed on the MASPOT population and Test panel DK. The phenotypic distributions of the three populations are depicted in Figure 4. For yield, it is noticeable that the MASPOT population contained individuals with very low phenotypic values, which can be explained by the fact that this was the only population where no selection had occurred. However, the MASPOT individuals also seem to have had some of the highest phenotypic values, while Test panel DK had a much narrower range in phenotypic values.

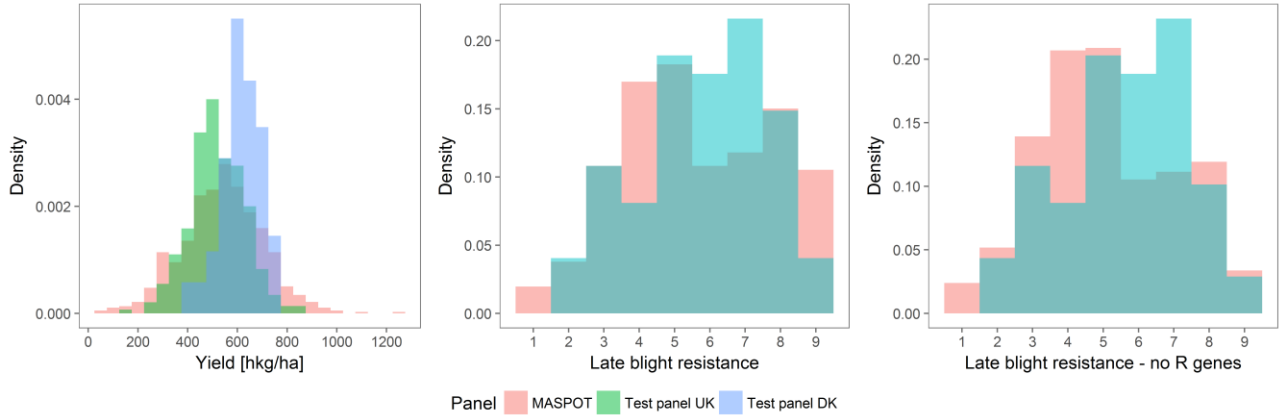

Figure 4 Density histograms of yield and late blight resistance (with and without R genes) phenotypes in the three populations. Yield was measured as hkg/ha, while late blight resistance was measured on a scale from 1 to 9, susceptible to resistant.

Low prediction correlations were obtained for yield of the MASPOT panel when using either the MASPOT population or Test panel DK as training population or the combined model (Table 3). Predictions for Test panel DK were moderate when using either the MASPOT population or the combined model for prediction, while predictions within the population were not as high. In all cases, when predicting Test panel DK, the bias was large, indicating a deflation in the scale relative to the observed phenotypes. As seen in Figure 5, the variation between the predicted values was small, and the high prediction correlations might merely be an indication of phenotypic variance in the observed values. Despite a much smaller GEBV bias within the MASPOT population, the prediction plots revealed the same effect. In fact, for all three populations, the predicted values ranged between 400 and 800, while the observed values for the MASPOT population were between 50 and 1400, and from 50 to nearly 1200 for Test panel UK. Still, there was a suggestion of some sort of correlation between predicted and observed values within Test panel UK, and yield predictions were quite good compared to the Danish populations, being almost twice as high as for within the MASPOT population. In contrast, predictions in Test panel UK using either of the Danish populations as training population were extremely poor. This could be an indication of a number of things, most likely related to the conversion of phenotypes from the measurement scale used for Test panel UK to the scale used for the Danish populations However, several different conversion methods were attempted, and in all cases, the prediction accuracy across populations was equally poor. It can thus not be ruled out that the lack of prediction ability is simply caused by genetic and/or environmental differences between the populations. Yield is strongly influenced by environmental and agricultural practices (Schönhals *et al.*, 2016). Furthermore, given the fact that yield is probably the most complex polygenic trait in potato, it is only logical that a vast number of markers, and a large number of individuals in the training population, is needed to capture all the genetic variation that is associated with yield. In this case, a high number of markers was utilised in the model (167,637), but due to the experimental setup with GBS, a large number of the individuals

have missing data for many markers, as seen in Paper 2. This led to much fewer markers being common for all individuals, and therefore much of the variation important for the trait was neglected. This could also explain the low prediction correlations in the MASPOT population and Test panel DK, as the yield predictions were generally poorer than predictions of starch content/dry matter and chipping quality (Paper 1 and 2). Another important factor was the low heritability of the trait. Narrow-sense heritability of yield within the MASPOT population was estimated to 20% from parent-offspring regression and 28% from genomic and phenotypic variances.

Table 3 Mean prediction correlations and biases for yield and late blight predictions performed with k-fold cross-validation with GBLUP over 10 repeats, using the populations separately and combined for modelling. The population used for training is listed horizontally while the predicted population is listed vertically. Bias is listed in brackets. The number of phenotypes available in each case is indicated with square brackets.

Prediction set / Training set	MASPOT	Test panel DK	Test panel UK	Combined
Yield				
MASPOT [753]	**0.25 [0.94]**	0.15 [0.69]	0.03 [0.18]	0.26 [0.83]
Test panel DK [69]	0.58 [1.64]	**0.33 [1.24]**	0.01 [0.04]	0.60 [1.39]
Test panel UK [290]	-0.03 [-0.18]	0.02 [0.36]	**0.48 [1.28]**	0.45 [1.45]
Late blight resistance				
MASPOT [712]	**0.56 [1.05]**	0.44 [1.25]	NA	0.56 [1.05]
Test panel DK [74]	0.37 [1.20]	**0.35 [1.18]**	NA	0.47 [1.29]
Late blight resistance – without R genes				
MASPOT [503]	**0.44 [1.09]**	0.22 [1.01]	NA	0.44 [1.05]
Test panel DK [69]	0.19 [0.88]	**0.49 [1.41]**	NA	0.45 [1.16]

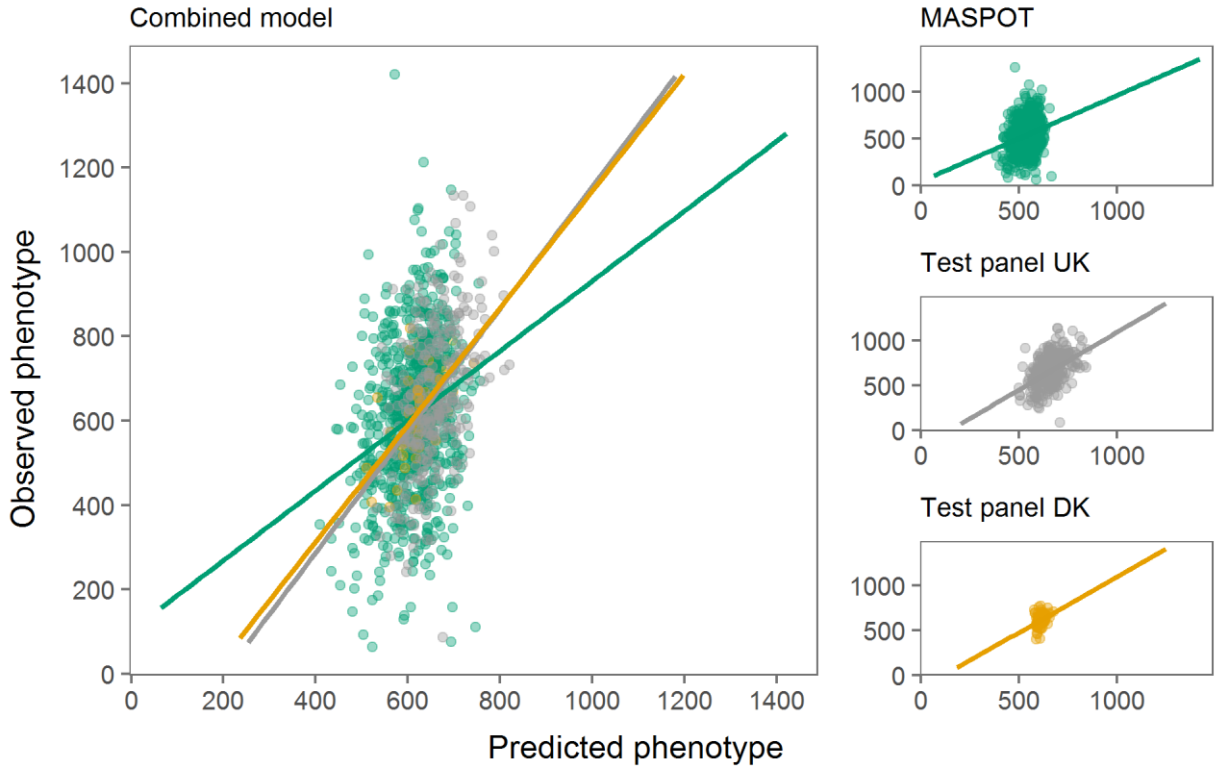

Figure 5 Prediction plots for yield with observed phenotype against predicted values. The big panel depicts predictions made with the combined model on the MASPOT population (green), Test panel UK (grey) and Test panel DK (yellow), while the right panels show predictions performed within each population.

Moderate prediction correlations were obtained for late blight resistance within the MASPOT population, while predictions within Test panel DK were slightly lower. Predictions for Test panel DK were the same whether the MASPOT population or Test panel DK was used, although, interestingly enough, prediction correlation was improved when combining the two populations. In theory, this would be expected, since a larger training population has the potential to capture a larger range of genetic diversity and provide more constraints on the variables in the prediction model. This would lead to better effect size estimates and better predictions. However, as seen in Paper 1 and Paper 2, merely increasing the training population size had practically no effect if the added individuals were derived from different populations, as least not for predictions of chipping quality and starch content/dry matter. These are highly quantitative and polygenic traits (van Eck, 2007), and although late blight resistance is also in part polygenic, it is also controlled by dominant R genes. As explained previously, 11 R genes have so far been identified and introgressed into cultivated potato from wild species (Umaerus and Umaerus, 1994). The chance of both populations having individuals with the same R genes is therefore large, and prediction of those highly resistant individuals is as a result simple, even if the populations are otherwise genetically or phenotypically diverse. However, when individuals with known R genes were removed, including some of the individuals exhibiting strongest late blight resistance, no improvement was seen in prediction accuracy between predictions within Test panel DK or when using both populations for predictions. This is likely because the remaining resistance is (mainly) quantitative and polygenic. A decrease in prediction accuracy was thus expected when removing the highly heritable resistance controlled by dominant R genes, and indeed, prediction correlations within the MASPOT population and across populations were lower compared to when using all individuals. Interestingly, there was a gain in prediction accuracy within Test panel DK, though the bias was also increased. As seen in the prediction plots in Figure 6 and Figure 7, the scale of predicted values was rather deflated relative to the observed values, especially for the predictions performed on individuals without known R genes. This can be adjusted for by simply transforming the scale when predicting within the same population, assuming that the same bias is common for the entire population. However, in terms of resistance to late blight, biased predictions can be critical, since a deflated scale means that the top performers will be underestimated, while the low performers will be overestimated, potentially leading to selection of the wrong candidates.

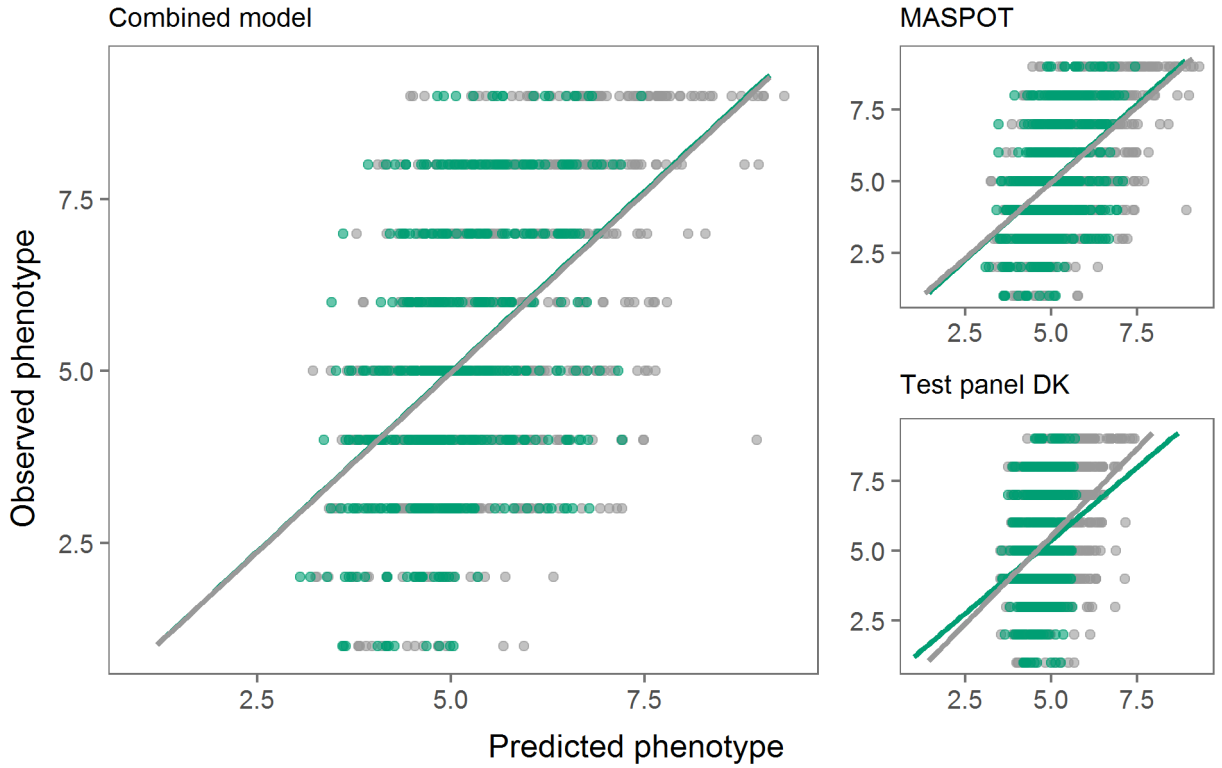

Figure 6 Late blight resistance predictions for MASPOT, using the combined model, the MASPOT population, or Test panel DK as training population. Predictions were made with all individuals (grey) and with only those individuals with no known R genes (green).

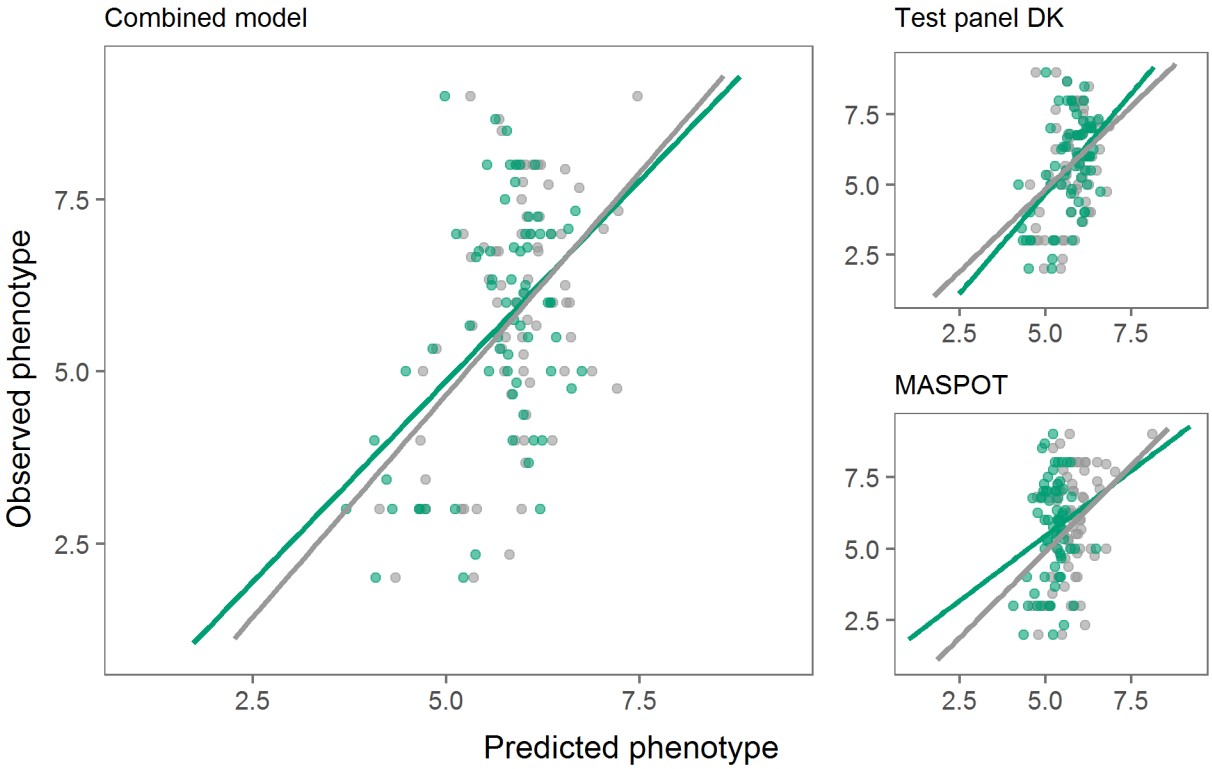

Figure 7 Late blight resistance predictions for Test panel DK, using the combined model, Test panel DK, or the MASPOT population as training population. Predictions were made with all individuals (grey) and with only those individuals with no known R genes (green).

Effect of marker number on prediction accuracy

In Paper 1 and Paper 2, respectively, 171,859 and 167,637 markers were used to generate genomic prediction models. Indeed, when making GS models, the goal is to capture as much of the genetic variation as possible, and in theory, a higher marker density will give improved predictions. However, as shown in Paper 2, the high number of markers was in fact redundant, and similar predictions could be made with significantly less markers. Figure 8 depicts prediction correlations obtained for predictions of chipping quality, dry matter and yield as a function of number of markers employed in the model. The predictions were conducted within the MASPOT population with 8-fold random cross-validation using GBLUP. As a starting point, the marker set of 167,637 markers from Paper 2 was used. Ten different selections were made for each marker amount, and 10 different cross-validation groupings were conducted for each selection. In the plot, prediction correlations for each sampling, i.e. 100 samplings per marker number, are plotted as boxplots to show the variation between the samplings.

For chipping quality and dry matter, prediction correlations increased as more SNPs were used in the model, reaching a plateau at around 10,000 markers, and at higher marker numbers the gain in prediction correlation was insignificant. As seen in the plot of prediction biases in Figure 9, predictions of each sampling fluctuated around a bias of 1 at low marker numbers. The biases became more stable around 10,000 markers, enhancing the same observation for the prediction correlation plot. Contrary to the prediction correlations, it appears that the GEBVs became more biased as the marker number increased, suggesting that the models become overfitted when the marker number is sufficient.

For yield, on the contrary, it appears that more markers were required to reach a stable level of prediction correlation. There was a slight indication of a plateau around 50,000 markers, although the prediction correlation might have reached a higher level if more markers were included. Bearing in mind that the GBS data used in this case contain a high number of missing data points, meaning that although a maximum number of markers was 167,637, these markers were not common for all individuals as a threshold of 50% missing data was allowed. As discussed previously, yield is a highly polygenic and complex trait, and it is therefore possible that predictions of yield are more sensitive to loss of marker density, as a high number of markers are required to capture the genetic diversity that is associated with the phenotypic diversity of yield. After all, even when using the entire marker set, yield prediction accuracies were significantly lower than those for chipping quality and dry matter. However, it has been found that non-additive genetic factors have predominant effect on tuber yield, which explains

the low prediction accuracies for yield as they are not considered in the models in this thesis (Plaisted *et al.*, 1962; Mendiburu and Peloquin, 1977).

For all three traits, the variation of prediction correlation and bias at each marker number between samplings was larger at lower marker numbers, signifying that a certain number of markers was needed for robustness of the model, although variations in yield predictions were significantly higher than for chipping quality and dry matter.

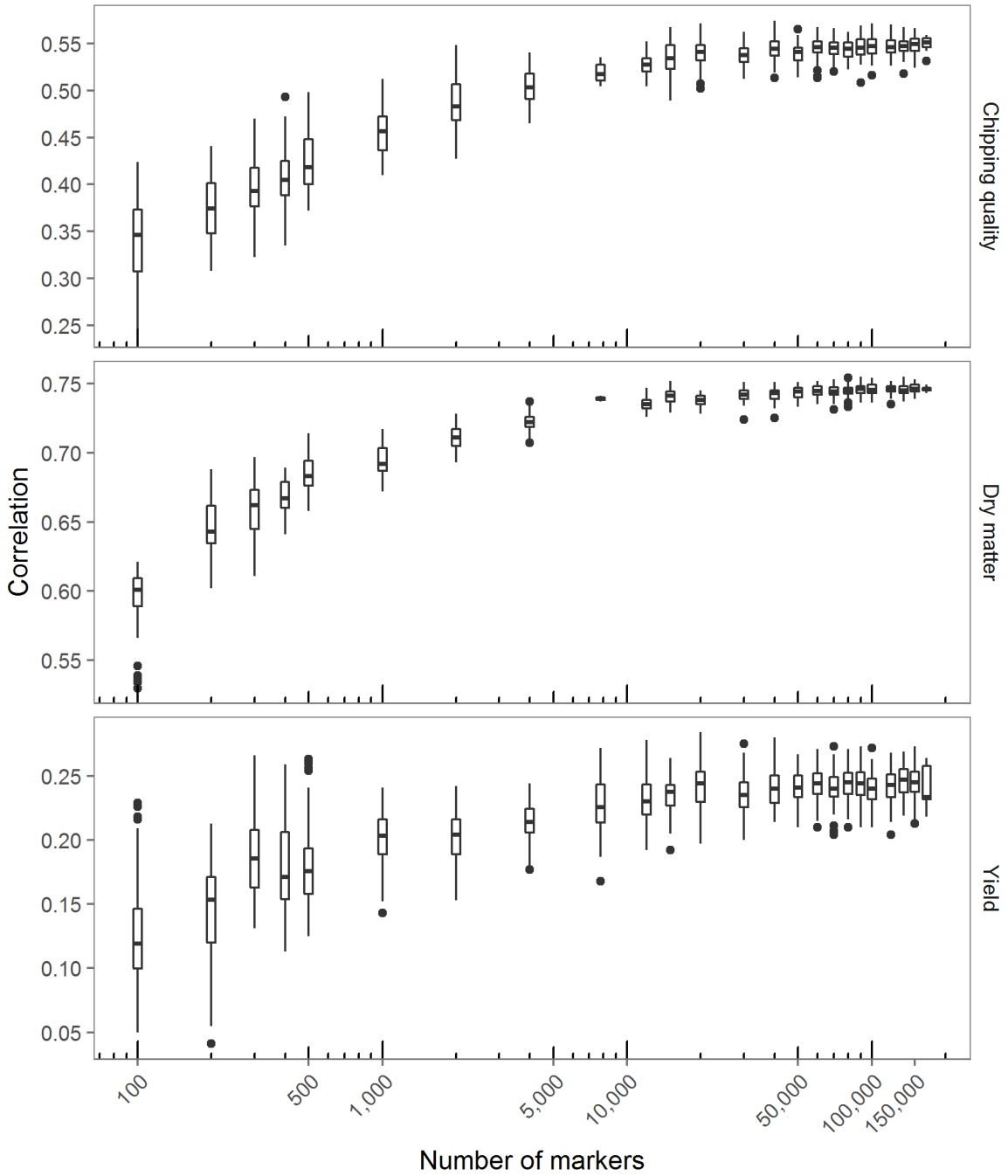

Figure 8 Prediction correlation between observed and predicted phenotypic values within the MASPOT population for chipping quality, dry matter content and yield over number of markers. The markers were selected randomly from the complete marker set of 167,637 markers, and each selection was repeated 10 times. 8-fold cross-validation was used, and predictions with each marker selection were repeated with 10 different cross-validation groupings. The top and bottom of the boxes correspond to first and third quartiles, while the centreline is the median, the whiskers extend to the lowest or the highest value that is within 1.5 x the inter-quartile range, and points represent outliers.

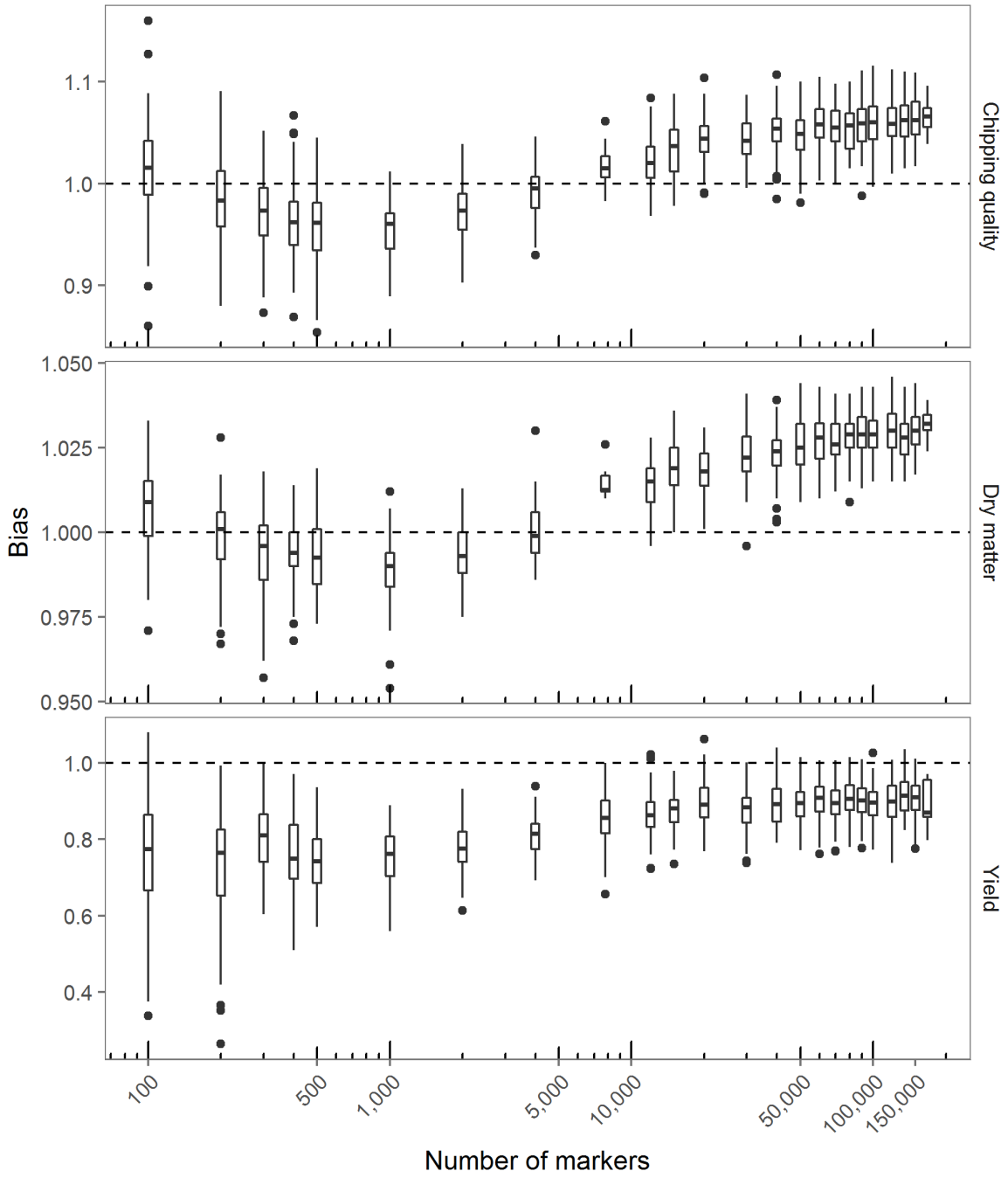

Figure 9 Bias of GEBVs estimated as the slope of the regression between observed and predicted phenotypic values within the MASPOT population for chipping quality, dry matter content and yield over number of markers. The markers were selected randomly from the complete marker set of 167,637 markers, and each selection was repeated 10 times. 8-fold cross-validation was used, and predictions with each marker selection were repeated with 10 different cross-validation groupings. The top and bottom of the boxes correspond to first and third quartiles, while the centreline is the median, the whiskers extend to the lowest or the highest value that is within 1.5 x the inter-quartile range, and points represent outliers.

Effect of training population size on prediction accuracy

In addition to marker number, the size of the training population has an impact on prediction accuracy of GS models. In Figure 10 and Figure 11, prediction correlations and prediction biases obtained for predictions of chipping quality and dry matter are plotted as a function of number of individuals. The predictions were conducted within the MASPOT population. First, all individuals with missing data for either of the two traits were removed for a fair comparison. 524 individuals remained, and from those, a number of individuals were selected from the population and used for prediction. The predictions were performed with k-fold random cross-validation using GBLUP. The cross-validation systems were constructed so that around 50 individuals were in each cross-validation group, except for the smallest group of 50 individuals, in which case 25 individuals were in each group. Ten different selections were made for each population size, and 10 different cross-validation groupings were conducted for each selection.

With a population of 50 individuals, poor predictions were obtained for both traits. Variations between selections were also large, especially for chipping quality predictions. Good predictions could be obtained, with predictions ranging up to above 0.6 for chipping quality and almost 0.80 for dry matter, but it depends on the selection of individuals. Thus, if by chance 50 genetically and phenotypically similar individuals were selected, good predictions could be obtained, but they might not make a good model for predicting individuals unlike themselves, given that they only cover a narrow range of genetic diversity. Predictions reached a high level around 300-350 individuals for both traits, both with regards to obtaining a high prediction correlation and a bias around 1 as well as less variations between samplings. Dry matter predictions seemed to reach a high level as soon as 200-250 individuals, although the variation between selections was considerable until reaching a more stable level around 300 individuals.

The difference in robustness between the two traits could be caused by a number of factors. For one, the heritability of a trait is significant to the prediction accuracy, and predictions of low-heritable traits require larger training populations. More individuals would give more observations per SNP allele and thus the SNP effects would be estimated more accurately. Heritability estimations from parent-offspring regressions for the MASPOT population estimated the heritability of chipping quality to be 74% and starch content 97% (see Results in Paper 1). Although the estimation of 97% heritability of starch content is rather high compared to other studies, it suggests that at least for the MASPOT population, starch content – and dry matter content – is a significantly more heritable trait than chipping quality, similar to what has been seen in other studies (Pereira *et al.*, 1994; Slater, Wilson, *et al.*, 2014). This explains the difference in prediction accuracies as well as the difference in robustness.

Another possible explanation is that there might be more non-additive genetic factors in play controlling the chipping quality trait. Indeed, this may very well be the case for chipping quality. Invertases have been shown to be important for chipping quality (Baldwin *et al.*, 2011; Schreiber *et al.*, 2014) and more than 20 loci encoding invertases exist in potato (Schreiber *et al.*, 2014). If multiple invertase loci can individually reduce the concentration of reducing sugars sufficiently to obtain high chipping quality, the effect of these loci is not additive, as each of them will have full effect.

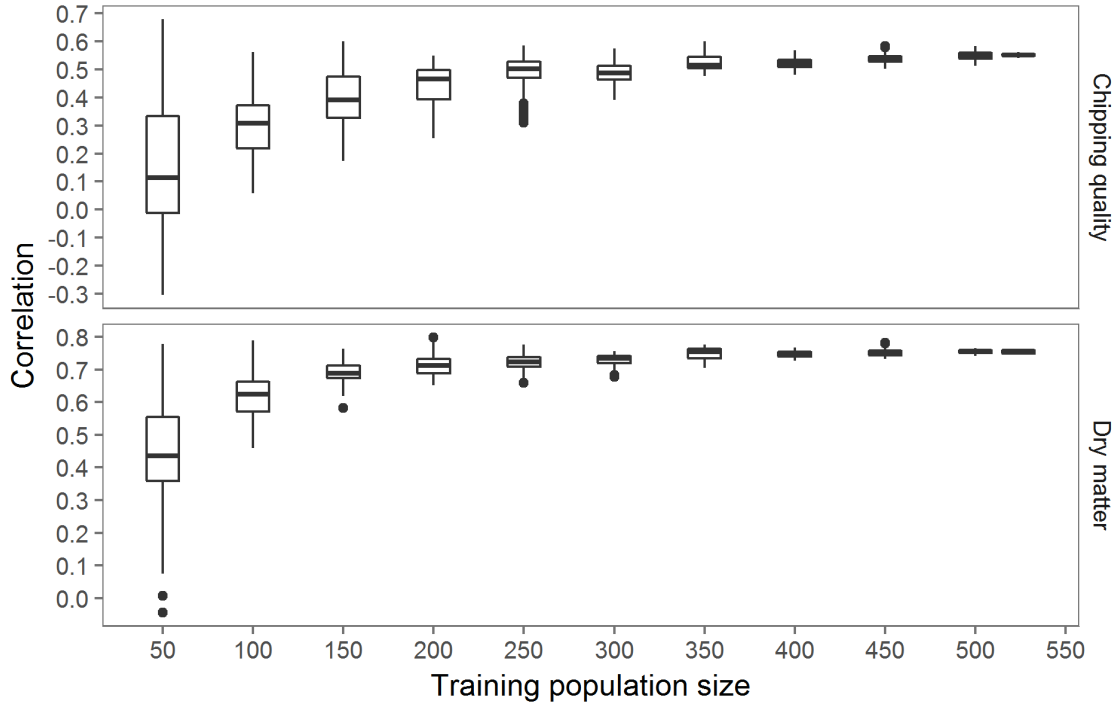

Figure 10 Boxplot showing prediction correlations of predictions within the MASPOT population at varying training population sizes between 50 and 524 individuals. All analyses were performed with 10 repeats. Predictions were performed using k-fold cross-validation systems with number of folds appropriate for the training population size. The top and bottom of the boxes correspond to first and third quartiles, while the centreline is the median, the whiskers extend to the lowest or the highest value that is within 1.5 x the inter-quartile range, and points represent outliers.

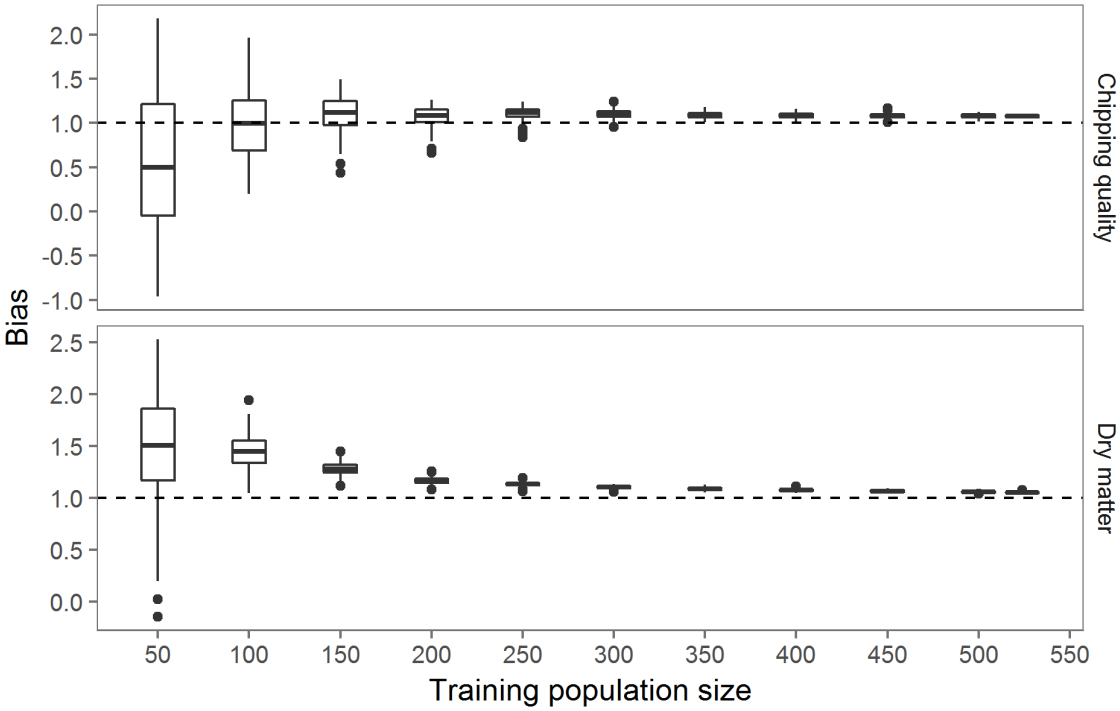

Figure 11 Boxplot showing prediction biases of predictions within the MASPOT population at varying training population sizes between 50 and 524 individuals. All analyses were performed with 10 repeats. Predictions were performed using k-fold cross-validation systems with number of folds appropriate for the training population size. The top and bottom of the boxes correspond to first and third quartiles, while the centreline is the median, the whiskers extend to the lowest or the highest value that is within 1.5 x the inter-quartile range, and points represent outliers.

Cross-validation systems

So far in the presented results in this thesis, random k-fold cross-validation systems have been utilised for generating prediction models. However, the choice of cross-validation system is important for the performance of prediction models. Disadvantages are connected to random k-fold cross-validation, in that it can give inflated estimates of the accuracy of genomic prediction, compared with what could actually be achieved in a breeding program. For example, in a diallel population like the MASPOT population, full and half sibs will be split across the training set and validation sets, which can lead to high accuracies resulting from predictions of effects of a limited number of very large chromosome segments due to their relatedness.

In order to investigate this effect, a leave-sibs-out cross-validation system was constructed, dividing the individuals in the MASPOT population into groups of full- and half-sibs. The 18 parents used for the MASPOT population were split into nine pairs, and the offspring were then divided into nine groups based on the parents, such that each group contained all offspring to one or both of the parents in the pair in question. Prediction correlations and biases for the leave-sibs-out cross-validated predictions are listed in Table 4 together with 8-fold cross-validated predictions for comparison (Paper 1).

Table 4 Leave-sibs-out cross-validated prediction correlations and biases for chipping quality and starch content obtained with BayesA, BayesC, and GBLUP. For comparison, the 8-fold random cross-validated prediction correlations from Paper 1 are included.

Trait / Cross-validation	BayesA		BayesC		GBLUP	
	Correlation	Bias	Correlation	Bias	Correlation	Bias
Chipping quality						
8-fold random cross-validation [524]	0.55	1.03	0.55	1.06	0.55	1.06
Leave-sibs-out cross-validation [524]	0.48	1.34	0.46	1.35	0.47	1.38
Starch content						
8-fold random cross-validation [755]	0.73	1.01	0.74	1.03	0.74	1.03
Leave-sibs-out cross-validation [755]	0.69	1.44	0.69	1.51	0.70	1.53

As expected, predictions obtained with the leave-sibs-out cross-validation system lead to lower prediction correlations, although the difference was small, especially for starch content. The biases were much larger, however, indicating that the scale was too small and that the top individuals were underestimated, while the lowest individuals were overestimated (Figure 12). When the individuals were separated based on their families, there would be bigger differences between each group than if they were grouped randomly. Hence, there was not as much range in the phenotypic values within each group, thus the scale of the predicted values would be smaller, and the bias larger. However, when predicting within the same population, biased predictions can be accounted for by transforming the scale, assuming that all individuals share the same degree of bias, although biased predictions cannot be compared across populations. The small difference in prediction accuracies for starch content between the two cross-validation systems suggests that the relationship between individuals did not play a large role in the prediction ability, while predictions of chipping quality might be slightly more influenced by family structures, in which case the more conservative cross-validation system based on pedigree would give more realistic prediction accuracies. Nonetheless, in a standard breeding program, where new breeding clones are generally made from crosses using a collection of the same elite cultivars, some kind of intermediate between random k-fold cross-validation and a leave-sibs-out cross-validation might be the most appropriate choice (Fè *et al.*, 2016).

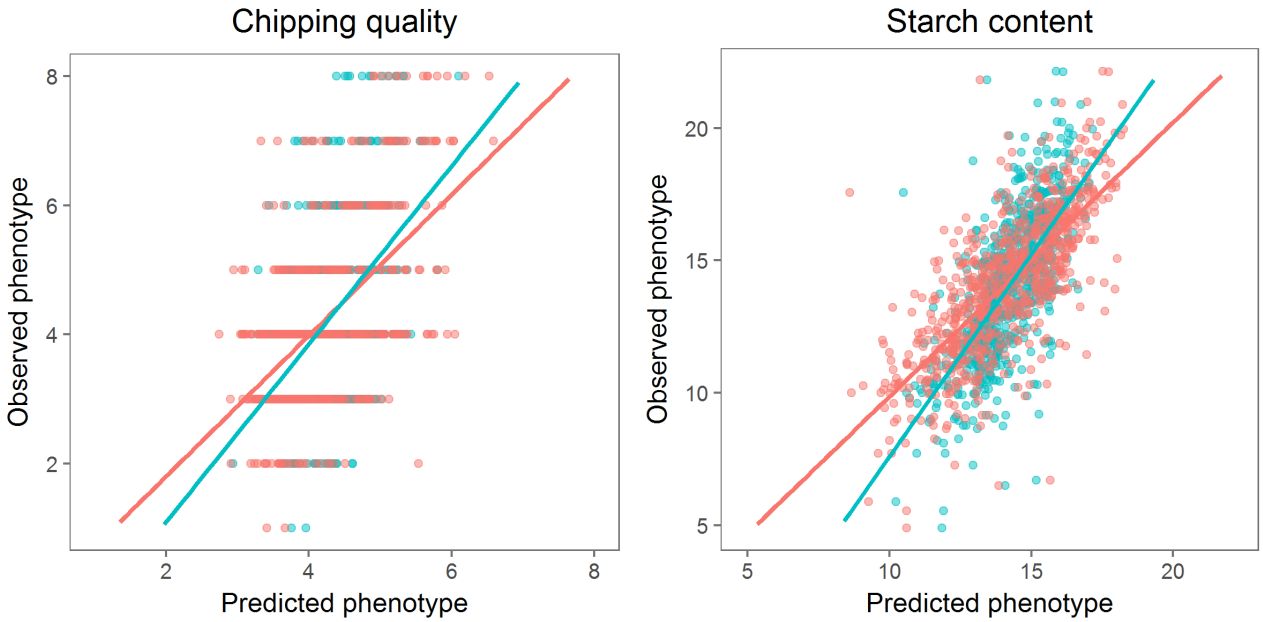

Figure 12 Prediction plots with predicted values against observed values for chipping quality (left) and starch content (right). Predictions were made with GBLUP within the MASPOT population with an 8-fold random cross-validation system (red) and a leave-sibs-out cross-validation system (blue).

Simulated bulk segregant analysis

A simulated BSA was performed on the GBS data and compared with results from a regular BSA study on dry matter. High and low performing individuals were selected from the MASPOT population in bulks of 96 each, though the "regular" BSA study utilised the entire MASPOT population of ~5,000 individuals. Figure 13 depicts the phenotypic distributions in each bulk from both studies, where it is clear that the bulks selected from the original MASPOT population did indeed include the most extreme individuals, while the bulks selected from the reduced MASPOT population were characterised by a prior selection and thus contained individuals of more average phenotypic values.

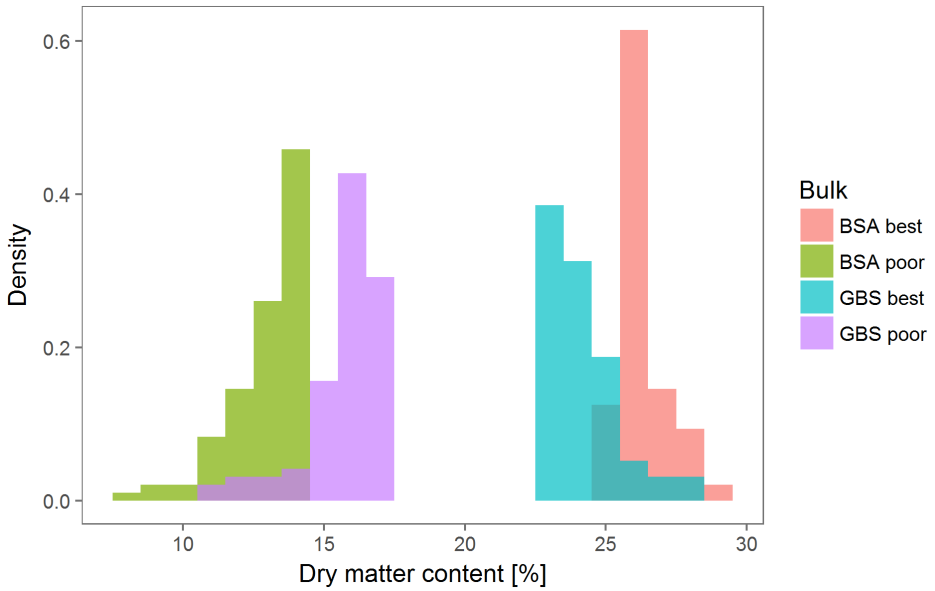

Figure 13 Density histogram of dry matter content distributions in the bulks selected for BSA in the original MASPOT population (BSA best and BSA poor) and in the reduced MASPOT population among GBS data (GBS best and GBS poor).

The BSA results for each analysis are depicted in the Circos plot in Figure 14 (Krzywinski *et al.*, 2009). In the plot, the values have been normalised for each chromosome. There were a number of significant regions identified with the original BSA results. In fact, chromosome 4 and 11 were the only chromosomes, where neither the best vs poor or poor vs best analyses found any significant regions. The GBS results revealed only a single significant region, which was on chromosome 1 (position 84,181,391-87,589,467). ADP-glucose pyrophosphorylase S (AGPaseS-1/2) has been mapped to this region (position 86,092,270-86,097,270) (Schreiber *et al.*, 2014), which is known to be one of the key enzymes controlling starch metabolism (Müller-Röber *et al.*, 1990), and it is one of three expressed genes encoding the large (L) subunit of glucose-1-phosphate adenylyltransferase.

Several peaks are common between the GBS data and the BSA data, even though the peaks were not significant in the GBS data. For example, the significant peak on chromosome 10 (position 53,245,546-56,342,423) found in best vs poor is also clearly visible in the GBS data. The same can be said for the peaks at chromosome 3 (positions 488,071-1,871,345 and 47,415,336-55,818,308), chromosome 6 (position 50,327,466-55,194,532), chromosome 8 (position 47,693,578-50,320,151), chromosome 9 (position 2,781,248-7,627,810), and chromosome 12 (positions 89,451-5,135,000 and 52,111,498-56,448,452).

The fact that only a single region was found to be significant for the GBS data suggests that this method of doing BSA with GBS data does not have as much power as a regular BSA. Bonferroni was also attempted to determine significance levels (data not shown), resulting in too many significant regions (>50%), while FDR on the genome-wide scale (rather than chromosome-wise) did not give any significant regions. The loss of power could be caused by a number of things. The clear difference in extreme phenotypic values between the bulks seen in Figure 13 would likely have a major influence on the power of determining relevant QTL. Additionally, the regular BSA was executed on approximately 9 million markers, while the GBS data consisted of 171,859 markers. Furthermore, statistical power could be increased by adopting a less conservative estimate of the null-distribution used to determine the significance threshold. In this analysis, all observed values were used to model the null distribution, essentially assuming that no QTL were present. This may be over-conservative. Alternatively, obvious peak regions may be excluded, assuming that they represent QTL. However, such an approach increases the risk of false positives and may be difficult to control and apply consistently. One possible consistent algorithm could be to exclude values that are three standard deviations away from the mean, and calculate the null distribution from the remaining values. Nonetheless, it is reassuring that peaks in the two methods were co-occurring.

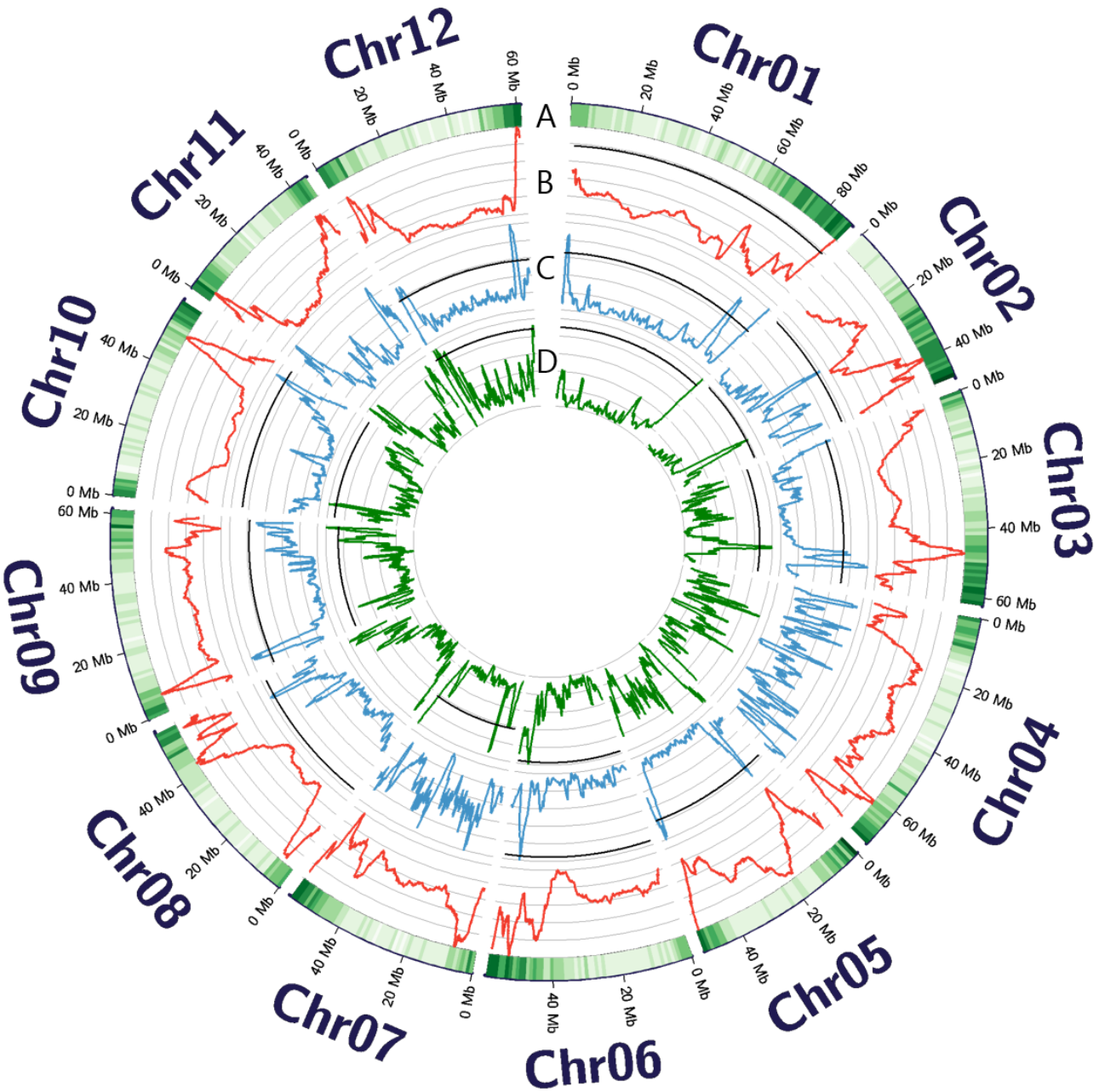

Figure 14 Circos plot with BSA results for dry matter content. A) Heat map of gene density ranging between 0 and 150 genes/Mb. B) Simulated BSA performed on GBS data. C) Chi-square test on best vs poor bulk. D) Chi-square test on poor vs best bulk. Black horizontal lines indicate FDR significance threshold determined for each chromosome. All values have been normalised between 0 and 1 for each chromosome.

In order to validate the significant QTL determined with BSA, markers in those regions were selected from the GBS data and genomic prediction models were constructed with GBLUP. Predictions were conducted within the MASPOT population (subset of 755) with 8-fold cross-validation. A total of 31,071 SNPs were selected. For comparison, 10 different sets of the same number of SNPs were selected randomly from the data set and predictions were carried out with the same conditions. Average cross-validated prediction correlation of 0.72 was obtained for dry matter predictions using the QTL selected from the BSA, with a bias of 1.02, similar to when predictions were performed with all 171,859 markers (see Paper 1). However, slightly higher prediction correlations were obtained when conducting predictions with 31,071 randomly selected SNPs, ranging between 0.74 and 0.75 for the 10 different marker sets and bias of 1.03-1.04. The prediction plot showing predictions from both methods is depicted in Figure 15. These results suggest that selection of trait-specific markers was irrelevant for prediction models, as good predictions could be obtained using unselected genome-wide markers.

However, this could also be explained by the fact that the selected markers (31,071) were sufficient to obtain a good prediction model because the linkage from one marker to the trait is adequate anywhere on the genome. As seen in a previous paragraph (Effect of marker number on prediction accuracy, page 36), around 10,000 markers are needed for good predictions in both dry matter and chipping quality. It would be interesting to see if bigger difference in prediction accuracies occurred if the selected markers had been less than 10,000. Of particular interest is whether phasing the SNPs to depict haplotype markers (combination of SNPs) that are more accurate representatives of the allelic structure of sites, makes a significant improvement of prediction models. However, this is not possible using the existing datasets.
The results are interesting since they exemplify that GBS data can be used for multiple purposes. Granted, the advantage of BSA is the reduced genotyping costs, since instead of multiple individuals, only two bulks are sequenced. However, BSA is also trait specific, while the same GBS data can be used to produce good prediction models for multiple traits, because the same genotyping data set can be regressed to different phenotyping datasets.

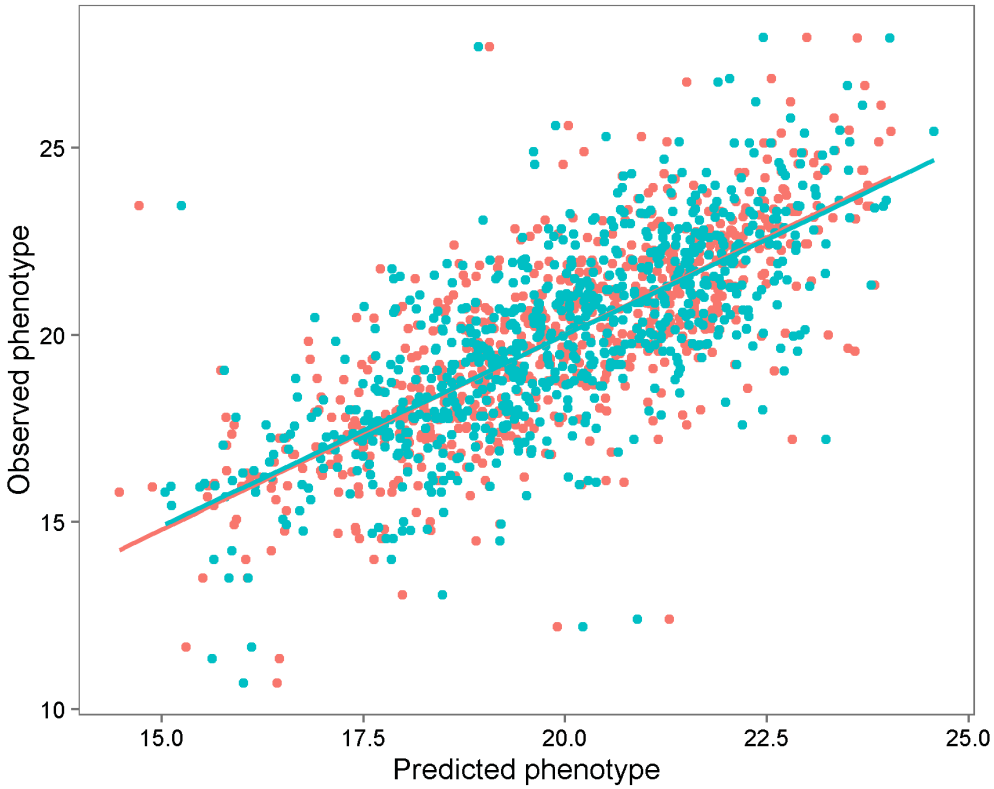

Figure 15 Prediction plot of observed phenotypes against predicted phenotypes for the simulated BSA study. Red: Mean value of predictions made from 10 sets of 31,071 randomly selected markers. Blue: Predictions made from 31,071 markers specifically selected as significant SNPs from BSA.

Genome wide association analysis

A GWAS was performed on the GBS data for the MASPOT population for dry matter content. Figure 16 depicts a Circos plot with the GWAS results. A total of 560 SNPs were found to be significant determined with FDR. A few significant SNPs were identified on chromosome 1 and 7, but the main part was found on chromosome 10. A number of invertases believed to be associated with a number of tuber quality traits are located on chromosome 10 (Schreiber *et al.*, 2014), including apoplastic invertase (*Inv-ap-b*) (Li *et al.*, 2008), which has been found to be associated with tuber starch content and chipping quality.

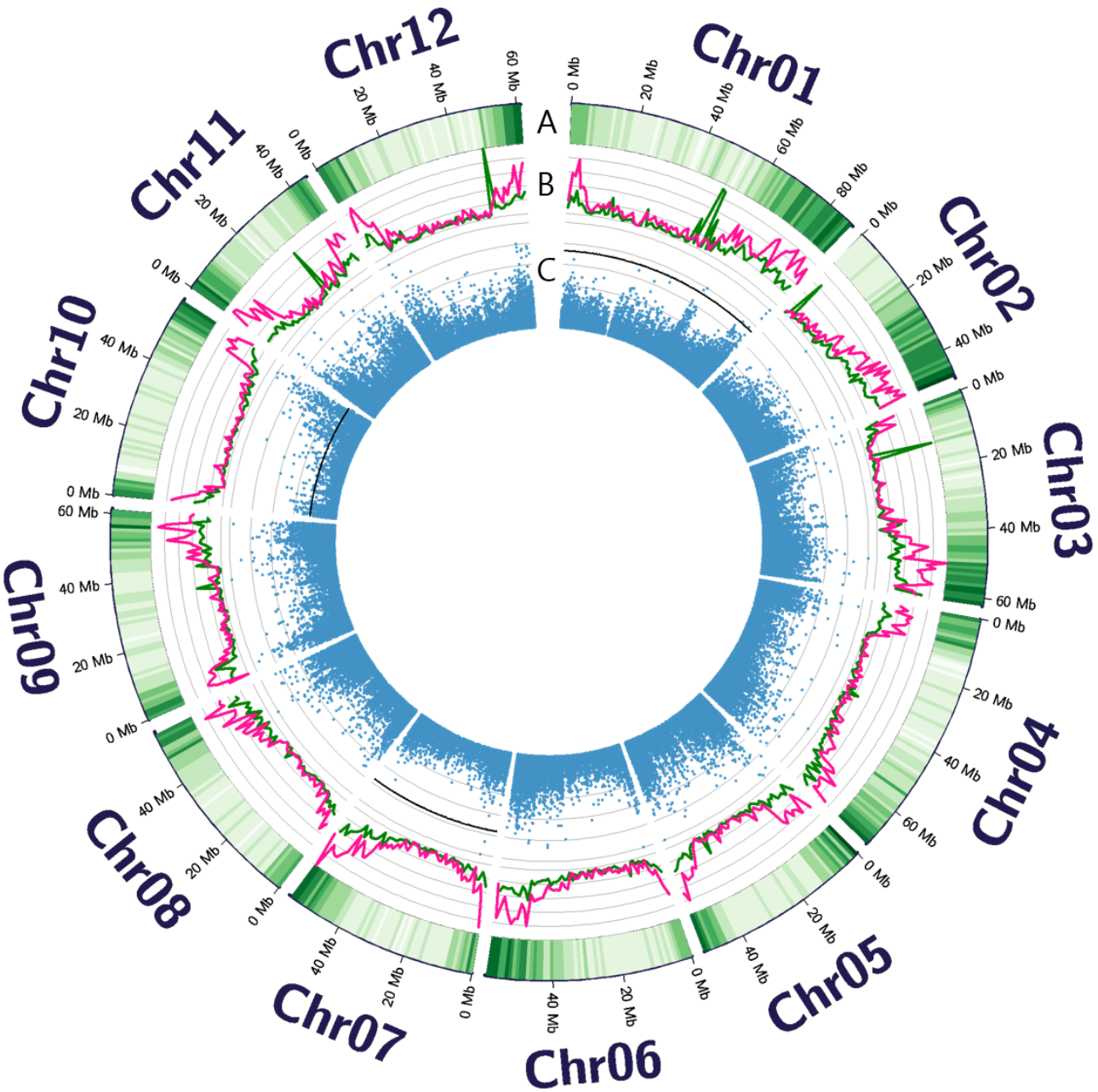

Figure 16 GWAS analysis of GBS data for dry matter content. A) Heat map of gene density ranging between 0 and 150 genes/Mb. B) Average coverage (green) and distribution of filtered markers (pink) in 1 Mb bins, normalised to the highest genome wide value. C) GWAS results as -log10 transformed p-values normalised between 0 and 1. Black bars indicate FDR significance threshold for each chromosome.

An important discussion about GS is whether the genome-wide prediction can keep up with predictions conducted with markers selected specifically for the purpose. Similarly to the analysis described in the previous paragraph concerning BSA, the 560 significant SNPs found with GWAS were selected from the GBS data and used to make prediction models. In addition, 10 different sets of 560 randomly selected SNPs were also used to make prediction models. These SNPs were selected based on having a correlation with the dry matter trait, however, for comparison, predictions were performed on both dry matter and chipping quality. High prediction correlation at 0.75 was obtained for dry matter content using the SNPs identified in the GWAS (Table 5). Similar prediction correlations were obtained when using the randomly selected SNPs, ranging between 0.69 and 0.72 for the 10 different marker selections. However, in the previous paragraph we saw that high correlations of dry matter predictions could be obtained even with a low number of markers, while the variation between each selection of markers can be significant for both the correlation and the bias. In this case, even though the

prediction correlations were similar, the results suggest that the 560 SNPs selected based on the GWAS results give slightly more powerful prediction models for predicting dry matter content.
Both methods also gave similar predictions when predicting chipping quality. Correlation of 0.44 was obtained with the GWAS selected SNPs, while correlations obtained for the randomly selected SNPs ranged between 0.41-0.47 for the 10 sets. Biases were also similar (1.04 and 0.98-1.02). It should be noted once again that the GWAS, and therefore the selection of markers, was performed on the dry matter trait and not chipping quality, but since tuber starch and sugar content are inversely related (Li *et al.*, 2013; Schreiber *et al.*, 2014), it is not surprising that the model, constructed from SNPs selected specifically for dry matter content, also performed reasonably when predicting chipping quality. It is thus possible that markers selected for dry matter or starch content predictions can be utilised for chipping quality as well.

Table 5 Prediction correlations and biases for predictions performed with 560 SNPs, either selected based on GWAS results or selected randomly.

Trait / Parameter	SNPs selected from GWAS results	Randomly selected
Dry matter		
Correlation	0.75	0.69-0.72
Bias	1.01	1.00-1.01
Chipping quality		
Correlation	0.44	0.41-0.47
Bias	1.04	0.98-1.02

Conclusions and future perspectives

Genomic prediction models were constructed for important traits in tetraploid potato within and across three populations. High prediction accuracies were obtained within each population for dry matter and starch content, while prediction accuracies for chipping quality were generally slightly lower. Predictions of tuber yield were significantly poorer, due to low heritability and a high proportion of non-additive genetic effects. Interestingly, prediction accuracies for both yield and chipping quality were significantly higher within one of the populations, coming from another breeding station than the other two, demonstrating the vast genetic and phenotypic diversity possible between tetraploid potato populations. Predictions of late blight resistance were moderate, and predictions were slightly poorer when excluding individuals with known R genes from the model, exemplifying the difference between dominant and quantitative resistance.

Predicting performance of individuals across breeds and genotypes is of great interest for breeders. However, low or moderate predictions were generally obtained across populations, similar to what has been observed in other plant species and animals. In all cases, prediction bias was large, i.e. the scale of predicted values was deflated, or in a few cases, inflated, relative to the observed phenotypic values. Predictions were especially poor for chipping quality and to a greater extent for yield across populations where the trait heritability and the within-population predictions were significantly different.

When expanding the training population to include all three populations used in this study, no gain in prediction was observed compared to predictions within each population. However, models constructed with this combined training population led to maximal prediction accuracy for all populations simultaneously, and could therefore be applied to all populations. This suggests that it is indeed possible to obtain a general potato prediction model if all relevant genotypes are included in the model, which is of great comfort, since this would make the GS models highly flexible.

An abundance of markers were used in the prediction models for this thesis, which became apparent when reducing the number markedly did not cause any significant prediction losses. No significant differences were observed between reduction approaches whether markers were chosen randomly or specifically selected based either on GWAS results, BSA results, or with the purpose of reducing the number of missing data. This confirms the assumption what GS is based on, which is that all QTL are in linkage disequilibrium with at least one marker, and that genome-wide markers are able to capture all (or most) of the genetic variance because of this.
Predictions of yield showed a higher sensitivity to marker reduction, possibly due to the high complexity and polygenic nature of the trait, although non-additive effects were also a major complication for yield predictions in general. Better predictions of yield could possibly be obtained by implementing non-parametric models able to model non-additive effects, but this was beyond the scope of this thesis.

Genotyping was performed with GBS, which proved itself a relatively simple and flexible genotyping method. Some complications are connected to GBS, most significantly the amount of missing data, particularly due to mutations in restriction sites. This especially has an impact when expanding a training population. Several imputation methods are available for imputation of missing GBS data, though as shown in this thesis, a lower marker number is sufficient for good predictions, and it is therefore unclear how important this problem is. Nevertheless, GBS proved itself a versatile genotyping method and showed potential for promising application methods. Besides GS, GBS data could be used for selecting significant SNPs with GWAS, and GBS is thus an excellent technique for identifying markers, e.g. for MAS. GWAS with GBS data can therefore supplement – or replace – the use of BSA, as GBS data can be used for marker selection for multiple traits simultaneously, contrary to BSA.

An excellent alternative to GBS is to produce a SNP chip with 10,000 genome-wide markers, although a higher marker number might be required for complex traits such as tuber yield. Possible reluctances against preproduced SNP chips are the restrictions and inflexibility surrounding it and the choice of markers. However, as shown in this thesis, even a low number of randomly selected markers is sufficient to produce good and robust prediction models.

Although an expanded training population displaying broad application across breeds was established in this thesis, it would be interesting to create an even larger population, encompassing a wide diversity of genotypes. An obtainable way to

accomplish this is to utilise the vast amounts of historical phenotypic data that has been gathered through decades at various breeding stations while genotyping the material that is in storage. This would require efficient pre-processing of phenotypic data, as especially environmental variations over time has significant impacts on phenotypic values. However, incorporation of a wide variety of genotypes in genomic prediction models would potentially result in more robust predictions with a broad application across breeds. Moreover, this would reduce start-up costs that are involved in establishing and phenotyping new breeding populations.

One of the major obstacles for implementing GS in existing breeding programmes is in fact the high investment costs required. However, multiple studies have shown that GS can potentially enhance genetic gain significantly, while consuming less cost per unit time compared to traditional breeding. Genetic gain in potato has been extremely slow, and no significant improvements in the yield potential of potato cultivars has been achieved over the last century. Our results demonstrate that GS is a promising breeding strategy for tetraploid potato.

References

Alexandratos, N. and Bruinsma, J. (2012) 'World agriculture: towards 2015/2030: an FAO perspective', *Land Use Policy*, 20(4), p. 375. doi: 10.1016/S0264-8377(03)00047-4.

Arruda, M. P., Brown, P. J., Lipka, A. E., Krill, A. M., Thurber, C. and Kolb, F. L. (2015) 'Genomic Selection for Predicting Fusarium Head Blight Resistance in a Wheat Breeding Program', *The Plant Genome*, 8(november), pp. 1–12. doi: 10.3835/plantgenome2015.01.0003.

Ashraf, B. H., Byrne, S., Fé, D., Czaban, A., Asp, T., Pedersen, M. G., Lenk, I., Roulund, N., Didion, T., Jensen, C. S., Jensen, J. and Janss, L. L. (2016) 'Estimating genomic heritabilities at the level of family-pool samples of perennial ryegrass using genotyping-by-sequencing', *Theoretical and Applied Genetics*. Springer Berlin Heidelberg, 129(1), pp. 45–52. doi: 10.1007/s00122-015-2607-9.

Baldwin, S. J., Dodds, K. G., Auvray, B., Genet, R. A., Macknight, R. C. and Jacobs, J. M. E. (2011) 'Association mapping of cold-induced sweetening in potato using historical phenotypic data', *Annals of Applied Biology*, 158(3), pp. 248–256. doi: 10.1111/j.1744-7348.2011.00459.x.

Barrell, P. J., Meiyalaghan, S., Jacobs, J. M. E. and Conner, A. J. (2013) 'Applications of biotechnology and genomics in potato improvement', *Plant Biotechnology Journal*, 11(8), pp. 907–920. doi: 10.1111/pbi.12099.

Birch, P. R. J., Bryan, G., Fenton, B., Gilroy, E. M., Hein, I., Jones, J. T., Prashar, A., Taylor, M. A., Torrance, L. and Toth, I. K. (2012) 'Crops that feed the world 8: Potato: Are the trends of increased global production sustainable?', *Food Security*, 4(4), pp. 477–508. doi: 10.1007/s12571-012-0220-1.

Boichard, D., Guillaume, F., Baur, A., Croiseau, P., Rossignol, M. N., Boscher, M. Y., Druet, T., Genestout, L., Colleau, J. J., Journaux, L., Ducrocq, V. and Fritz, S. (2012) 'Genomic selection in French dairy cattle', *Animal Production Science*, 52(2–3), pp. 115–120. doi: 10.1071/AN11119.

Bonierbale, M. W., Plaisted, R. L. and Tanksley, S. D. (1993) 'A test of the maximum heterozygosity hypothesis using molecular markers in tetraploid potatoes', *Theoretical and Applied Genetics (TAG)*, 86, pp. 481–491.

Borlaug, N. E. (2002) 'Feeding a world of 10 billion people: The miracle ahead', *In Vitro Cellular & Developmental Biology - Plant*, 38(2), pp. 221–228. doi: 10.1079/IVP2001279.

Bourke, A. (1993) *'The visitation of god?' The potato and the great Irish famine.* Dublin: Lilliput Press Ltd.

Bradshaw, J. E. (2006) 'Genetics of Agrihorticultural Traits', in Gopal, J. and Khurana, S. M. P. (eds) *Handbook of Potato Production, Improvement, and Postharvest Management*. Food Products Press, pp. 41–75.

Bradshaw, J. E. (2007) 'Potato-Breeding Strategy', in Vreugdenhil, D., Bradshaw, J., Gebhardt, C., Govers, F., MacKerron, D. K. L., Taylor, M. A., and Ross, H. A. (eds) *Potato Biology and Biotechnology: Advances and Perspectives*. Amsterdam: Elsevier, pp. 157–174.

Bradshaw, J. E., Dale, M. F. B. and Mackay, G. R. (2003) 'Use of mid-parent values and progeny tests to increase the efficiency of potato breeding for combined processing quality and disease and pest resistance.', *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik*, 107(1), pp. 36–42. doi: 10.1007/s00122-003-1219-y.

Bradshaw, J. E., Dale, M. F. B. and Mackay, G. R. (2009) 'Improving the yield, processing quality and disease and pest resistance of potatoes by genotypic recurrent selection', *Euphytica*, 170(1), pp. 215–227. doi: 10.1007/s10681-009-9925-4.

Brown, J., Mackay, G. R., Bain, H., Griffith, D. W. and Allison, M. J. (1990) 'The processing potential of tubers of the cultivated potato,Solanum tuberosum L., after storage at low temperatures. 2. Sugar concentration', *Potato Research*, 33, pp. 219–227. doi: 10.1007/BF02358449.

Burlingame, B., Mouillé, B. and Charrondière, R. (2009) 'Nutrients, bioactive non-nutrients and anti-nutrients in potatoes', *Journal of Food Composition and Analysis*, 22(6), pp. 494–502. doi: 10.1016/j.jfca.2009.09.001.

Calus, M. P. L., Meuwissen, T. H. E., De Roos, A. P. W. and Veerkamp, R. F. (2008) 'Accuracy of genomic selection using different methods to define haplotypes', *Genetics*, 178(1), pp. 553–561. doi: 10.1534/genetics.107.080838.

Chen, Q., Kawchuk, L. M., Lynch, D. R., Goettel, M. S. and Fujimoto, D. K. (2003) 'Identification of late blight, Colorado potato beetle, and blackleg resistance in three Mexican and two South American wild 2x (1EBN)Solanum species', *American Journal of Potato Research*, 80(1), pp. 9–19. doi: 10.1007/BF02854552.

Chen, X., Salamini, F. and Gebhardt, C. (2001) 'A potato molecular-function map for carbohydrate metabolism and transport', *Theoretical and Applied Genetics*, 102(2–3), pp. 284–295. doi: 10.1007/s001220051645.

Clasen, B. M., Stoddard, T. J., Luo, S., Demorest, Z. L., Li, J., Cedrone, F., Tibebu, R., Davison, S., Ray, E. E., Daulhac, A., Coffman, A., Yabandith, A., Retterath, A., Haun, W., Baltes, N. J., Mathis, L., Voytas, D. F. and Zhang, F. (2016) 'Improving cold storage and processing traits in potato through targeted gene knockout', *Plant Biotechnology Journal*, 14(1), pp. 169–176. doi: 10.1111/pbi.12370.

Clifford, D. and McCullagh, P. (2006) 'The regress function, R News 6:2, 6-10'.

Clifford, D. and McCullagh, P. (2014) 'The regress package R package version 1.3-14'.

Collins, A., Milbourne, D., Ramsay, L., Meyer, R., Chatot-Balandras, C., Oberhagemann, P., De Jong, W., Gebhardt, C., Bonnel, E. and Waugh, R. (1999) 'QTL for field resistance to late blight in potato are strongly correlated with maturity and

vigour', *Molecular Breeding*, pp. 387–398. doi: 10.1023/A:1009601427062.
Crossa, J., De Los Campos, G., Pérez, P., Gianola, D., Burgueño, J., Araus, J. L., Makumbi, D., Singh, R. P., Dreisigacker, S., Yan, J., Arief, V., Banziger, M. and Braun, H. J. (2010) 'Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers', *Genetics*, 186(2), pp. 713–724. doi: 10.1534/genetics.110.118521.
Cunningham, C. E. and Stevenson, F. J. (1963) 'Inheritance of factors affecting potato chip color and their association with specific gravity', *American Potato Journal*, 40(8), pp. 253–265. doi: 10.1007/BF02850325.
D'Hoop, B. B., Paulo, M. J., Mank, R. A., Van Eck, H. J. and Van Eeuwijk, F. A. (2008) 'Association mapping of quality traits in potato (Solanum tuberosum L.)', *Euphytica*, 161(1–2), pp. 47–60. doi: 10.1007/s10681-007-9565-5.
Daetwyler, H. D., Hickey, J. M., Henshall, J. M., Dominik, S., Gredler, B., Van Der Werf, J. H. J. and Hayes, B. J. (2010) 'Accuracy of estimated genomic breeding values for wool and meat traits in a multi-breed sheep population', *Animal Production Science*, 50(11–12), pp. 1004–1010. doi: 10.1071/AN10096.
Dale, M. E. B. and Mackay, G. R. (1994) 'Inheritance of table and processing quality', in Bradshaw, J. E. and Mackay, G. R. (eds) *Potato Genetics*, pp. 285–315.
Dekkers, J. C. M. and Hospital, F. (2002) 'The use of molecular genetics in the improvement of agricultural populations.', *Nature reviews. Genetics*, 3(1), pp. 22–32. doi: 10.1038/nrg701.
Dewey, M. (2017) 'metap: meta-analysis of significance values. R package version 0.8'.
Douches, D. S. and Freyre, R. (1994) 'Identification of genetic factors influencing chip color in diploid potato (Solanum spp.)', *American Potato Journal*, 71(9), pp. 581–590. doi: 10.1007/BF02851523.
Draffehn, A. M., Meller, S., Li, L. and Gebhardt, C. (2010) 'Natural diversity of potato (Solanum tuberosum) invertases.', *BMC plant biology*, 10, p. 271. doi: 10.1186/1471-2229-10-271.
van Eck, H. J. (2007) 'Genetics of Morphological and Tuber Traits', in Vreugdenhil, D., Bradshaw, J., Gebhardt, C., Govers, F., Taylor, M. A., MacKerron, D. K. L., and Ross, H. A. (eds) *Potato Biology and Biotechnology: Advances and Perspectives*. Amsterdam: Elsevier, pp. 91–115.
van Eck, H. J., van der Voort, J. R., Draaistra, J., van Zandvoort, P., van Enckevort, E., Segers, B., Peleman, J., Jacobsen, E., Helder, J. and Bakker, J. (1995) 'The inheritance and chromosomal localization of AFLP markers in a non-inbred potato offspring', *Molecular Breeding*, 1(4), pp. 397–410. doi: 10.1007/BF01248417.
El-Kharbotly, A., Leonards-Schippers, C., Huigen, D. J., Jacobsen, E., Pereira, A., Stiekema, W. J., Salamini, F. and Gebhardt, C. (1994) 'Segregation analysis and RFLP mapping of the R1 and R3 alleles conferring race-specific resistance to Phytophthora infestans in progeny of dihaploid potato parents', *MGG Molecular & General Genetics*, 242(6), pp. 749–754. doi: 10.1007/BF00283432.
El-Kharbotly, A., Palomino-Sánchez, C., Salamini, F., Jacobsen, E. and Gebhardt, C. (1996) 'R6 and R7 alleles of potato conferring race-specific resistance to Phytophthora infestans (Mont.) de Bary identified genetic loci clustering with the R3 locus on chromosome XI', *Theoretical and Applied Genetics*, 92(7), pp. 880–884. doi: 10.1007/s001220050206.
Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S. and Mitchell, S. E. (2011) 'A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species', *PLoS ONE*, 6(5). doi: 10.1371/journal.pone.0019379.
Ewing, E. E., Šimko, I., Smart, C. D., Bonierbale, M. W., Mizubuti, E. S. G., May, G. D. and Fry, W. E. (2000) 'Genetic mapping from field tests of qualitative and quantitative resistance to Phytophthora infestans in a population derived from Solanum tuberosum and Solanum berthaultii', *Molecular Breeding*, 6(1), pp. 25–36. doi: 10.1023/A:1009648408198.
FAOSTAT (2017) *Food and Agriculture Organization of the United Nations*. Available at: http://www.fao.org/faostat/ (Accessed: 8 March 2017).
Fè, D., Ashraf, B. H., Pedersen, M. G., Janss, L., Byrne, S., Roulund, N., Lenk, I., Didion, T., Asp, T., Jensen, C. S. and Jensen, J. (2016) 'Accuracy of Genomic Prediction in a Commercial Perennial Ryegrass Breeding Program', *The Plant Genome*, 9(3), pp. 1–22. doi: 10.3835/plantgenome2015.11.0110.
Feingold, S., Lloyd, J., Norero, N., Bonierbale, M. and Lorenzen, J. (2005) 'Mapping and characterization of new EST-derived microsatellites for potato (Solanum tuberosum L.)', *Theoretical and Applied Genetics*, 111(3), pp. 456–466. doi: 10.1007/s00122-005-2028-2.
Fischer, M., Schreiber, L., Colby, T., Kuckenberg, M., Tacke, E., Hofferbert, H.-R., Schmidt, J. and Gebhardt, C. (2013) 'Novel candidate genes influencing natural variation in potato tuber cold sweetening identified by comparative proteomics and association mapping', *BMC Plant Biology*, 13, p. 113. doi: 10.1186/1471-2229-13-113.
Forbes, G. A. and Landeo, J. A. (2006) 'Late blight', in Gopal, J. and Khurana, S. (eds) *Handbook of Potato Production, Improvement, and Postharvest Management*. Binghamton, NY, pp. 279–314.
Freyre, R. and Douches, D. S. (1994) 'Development of a model for marker-assisted selection of specific gravity in diploid potato across environments', *Crop Science*, 34(5), pp. 1361–1368. doi: 10.2135/cropsci1994.0011183X003400050040x.
Fry, W. E. and Goodwin, S. B. (1997) 'Resurgence of the Irish Potato Famine Fungus', *BioScience*, 47(6), pp. 363–371.
Gebhardt, C. (2013) 'Bridging the gap between genome analysis and precision breeding in potato', *Trends in Genetics*. Elsevier Ltd, 29(4), pp. 248–256. doi: 10.1016/j.tig.2012.11.006.
Gebhardt, C., Ritter, E., Debener, T., Schachtschabel, U., Walkemeier, B., Uhrig, H. and Salamini, F. (1989) 'RFLP analysis and linkage mapping in \emph{Solanum tuberosum}', *Theoretical and Applied Genetics*, 78, pp. 65–75. doi:

10.1007/bf00299755.
Gebhardt, C. and Valkonen, J. P. . (2001) 'Organization of genes controlling disease resistance in the potato genome', *Annual Review of Phytopathology*, 39(1), pp. 79–102. doi: 10.1146/annurev.arplant.54.031902.135035.
Ghislain, M., Spooner, D. M., Rodríguez, F., Villamón, F., Núñez, J., Vásquez, C., Waugh, R. and Bonierbale, M. (2004) 'Selection of highly informative and user-friendly microsatellites (SSRs) for genotyping of cultivated potato', *Theoretical and Applied Genetics*, 108(5), pp. 881–890. doi: 10.1007/s00122-003-1494-7.
Gianessi, L. and Williams, A. (2011) *Restrictions on Fungicide Use Causing Decline in Organic Potato Production in Europe*.
Gianola, D. (2013) 'Priors in whole-genome regression: The Bayesian alphabet returns', *Genetics*, 194(3), pp. 573–596. doi: 10.1534/genetics.113.151753.
Goddard, M. (2009) 'Genomic selection: Prediction of accuracy and maximisation of long term response', *Genetica*, 136(2), pp. 245–257. doi: 10.1007/s10709-008-9308-0.
Goddard, M. E. and Hayes, B. J. (2007) 'Genomic selection', *J. Anim. Breed. Genet.* , 124, pp. 323–330. doi: 10.1111/j.1439-0388.2007.00702.x.
Grattapaglia, D., Vilela Resende, M., Resende, M., Sansaloni, C., Petroli, C., Missiaggia, A., Takahashi, E., Zamprogno, K. and Kilian, A. (2011) 'Genomic Selection for growth traits in Eucalyptus: accuracy within and across breeding populations', *BMC Proceedings*, 5(Suppl 7), p. O16. doi: 10.1186/1753-6561-5-S7-O16.
Halterman, D., Guenthner, J., Collinge, S., Butler, N. and Douches, D. (2016) 'Biotech Potatoes in the 21st Century: 20 Years Since the First Biotech Potato', *American Journal of Potato Research*, 93(1), pp. 1–20. doi: 10.1007/s12230-015-9485-1.
Hayes, B., Bowman, P., Chamberlain, A. and Goddard, M. (2009) 'Invited review: genomic selection in dairy cattle: progress and challenges', *J Dairy Sci*, 92. doi: 10.3168/jds.2008-1646.
Hayes, B. J., Bowman, P. J., Chamberlain, A. J. and Goddard, M. E. (2009) 'Genomic selection in dairy cattle: progress and challenges', *Journal of Dairy Science*. Elsevier, 92(2), pp. 433–443. doi: 92/2/433 [pii];10.3168/jds.2008-1646 [doi].
He, J., Zhao, X., Laroche, A., Lu, Z.-X., Liu, H. and Li, Z. (2014) 'Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelerate plant breeding', *Frontiers in Plant Science*, 5(September), p. 484. doi: 10.3389/fpls.2014.00484.
Heffner, E. L., Sorrells, M. E. and Jannink, J. (2009) 'Genomic Selection for Crop Improvement', *Crop Science*, 49(1), pp. 1–12. doi: 10.2135/cropsci2008.08.0512.
Huang, S., Vleeshouwers, V. G. a a, Werij, J. S., Hutten, R. C. B., van Eck, H. J., Visser, R. G. F. and Jacobsen, E. (2004) 'The R3 resistance to Phytophthora infestans in potato is conferred by two closely linked R genes with distinct specificities.', *Molecular plant-microbe interactions : MPMI*, 17(4), pp. 428–435. doi: 10.1094/MPMI.2004.17.4.428.
Huang, S., Van Der Vossen, E. A. G., Kuang, H., Vleeshouwers, V. G. A. A., Zhang, N., Borm, T. J. A., Van Eck, H. J., Baker, B., Jacobsen, E. and Visser, R. G. F. (2005) 'Comparative genomics enabled the isolation of the R3a late blight resistance gene in potato', *Plant Journal*, 42(2), pp. 251–261. doi: 10.1111/j.1365-313X.2005.02365.x.
Isherwood, F. A. (1973) 'Starch-Sugar Interconversion in Solanum Tuberosum', *Phytochemistry*, 12(1965), pp. 2579–2591.
Jacobs, J. M. E., Van Eck, H. J., Arens, P., Verkerk-Bakker, B., te Lintel Hekkert, B., Bastiaanssen, H. J. M., El-Kharbotly, A., Pereira, A., Jacobsen, E. and Stiekema, W. J. (1995) 'A genetic map of potato (Solanum tuberosum) integrating molecular markers, including transposons, and classical markers', *Theoretical and Applied Genetics*, 91(2), pp. 289–300. doi: 10.1007/BF00220891.
Jarquín, D., Kocak, K., Posadas, L., Hyma, K., Jedlicka, J., Graef, G. and Lorenz, A. (2014) 'Genotyping by sequencing for genomic prediction in a soybean breeding population.', *BMC genomics*, 15(1), p. 740. doi: 10.1186/1471-2164-15-740.
Johnsen, H. Ø. (2015) *Marker assisted selection of dry matter content in potato breeding clones*. Aalborg University.
Kirkman, M. A. (2007) 'Global Markets for Processed Potato Products', in Vreugdenhil, D., Bradshaw, J., Gebhardt, C., Govers, F., MacKerron, D. K. L., Taylor, M. A., and Ross, H. A. (eds) *Potato Biology and Biotechnology: Advances and Perspectives*. 1st edn, pp. 27–44.
Krzywinski, M. I., Schein, J. E., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S. J. and Marra, M. A. (2009) 'Circos: An information aesthetic for comparative genomics', *Genome Research* . doi: 10.1101/gr.092759.109.
Leonards-Schippers, C., Gieffers, W., Salamini, F. and Gebhardt, C. (1992) 'The R1 gene conferring race-specific resistance to Phytophthora infestans in potato is located on potato chromosome V', *Molecular and General Genetics*, 233(1–2), pp. 278–283. doi: 10.1007/BF00587589.
Leonards-Schippers, C., Gieffers, W., Schafer-Pregl, R., Ritter, E., Knapp, S. J., Salamini, F. and Gebhardt, C. (1994) 'Quantitative resistance to Phytophthora infestans in potato: A case study for QTL mapping in an allogamous plant species', *Genetics*, 137(1), pp. 67–77.
Li, L., Paulo, M. J., Strahwald, J., Lübeck, J., Hofferbert, H. R., Tacke, E., Junghans, H., Wunder, J., Draffehn, A., Van Eeuwijk, F. and Gebhardt, C. (2008) 'Natural DNA variation at candidate loci is associated with potato chip color, tuber starch content, yield and starch yield', *Theoretical and Applied Genetics*, 116(8), pp. 1167–1181. doi: 10.1007/s00122-008-0746-y.
Li, L., Tacke, E., Hofferbert, H. R., Lübeck, J., Strahwald, J., Draffehn, A. M., Walkemeier, B. and Gebhardt, C. (2013) 'Validation of candidate gene markers for marker-assisted selection of potato cultivars with improved tuber quality', *Theoretical and Applied Genetics*, 126(4), pp. 1039–1052. doi: 10.1007/s00122-012-2035-z.

Li, X., Van Eck, H. J., Rouppe Van Der Voort, J. N. A. M., Huigen, D. J., Stam, P. and Jacobsen, E. (1998) 'Autotetraploids and genetic mapping using common AFLP markers: The R2 allele conferring resistance to Phytophthora infestans mapped on potato chromosome 4', *Theoretical and Applied Genetics*, 96(8), pp. 1121–1128. doi: 10.1007/s001220050847.
Lorenz, A. J., Chao, S., Asoro, F. G., Heffner, E. L., Hayashi, T., Iwata, H., Smith, K. P., Sorrells, M. E. and Jannink, J. L. (2011) *Genomic Selection in Plant Breeding. Knowledge and Prospects.*, *Advances in Agronomy*. doi: 10.1016/B978-0-12-385531-2.00002-5.
de los Campos, G., Hickey, J. M., Pong-Wong, R., Daetwyler, H. D. and Calus, M. P. L. (2013) 'Whole-genome regression and prediction methods applied to plant and animal breeding', *Genetics*, 193(2), pp. 327–345. doi: 10.1534/genetics.112.143313.
de los Campos, G. and Perez Rodriguez, P. (2015) 'BGLR: Bayesian Generalized Linear Regression'.
Luan, T., Woolliams, J. A., Lien, S., Kent, M., Svendsen, M. and Meuwissen, T. H. E. (2009) 'The accuracy of genomic selection in Norwegian red cattle assessed by cross-validation', *Genetics*, 183(3), pp. 1119–1126. doi: 10.1534/genetics.109.107391.
Lutaladio, N. and Castaldi, L. (2009) 'Potato: The hidden treasure', *Journal of Food Composition and Analysis*, 22(6), pp. 491–493. doi: 10.1016/j.jfca.2009.05.002.
Ma, Y., Reif, J. C., Jiang, Y., Wen, Z., Wang, D., Liu, Z., Guo, Y., Wei, S., Wang, S., Yang, C., Wang, H., Yang, C., Lu, W., Xu, R., Zhou, R., Wang, R., Sun, Z., Chen, H., Zhang, W., Wu, J., Hu, G., Liu, C., Luan, X., Fu, Y., Guo, T., Han, T., Zhang, M., Sun, B., Zhang, L., Chen, W., Wu, C., Sun, S., Yuan, B., Zhou, X., Han, D., Yan, H., Li, W. and Qiu, L. (2016) 'Potential of marker selection to increase prediction accuracy of genomic selection in soybean (Glycine max L.)', *Molecular Breeding*. Springer Netherlands, 36(8), pp. 1–10. doi: 10.1007/s11032-016-0504-9.
Magwene, P. M., Willis, J. H. and Kelly, J. K. (2011) 'The statistics of bulk segregant analysis using next generation sequencing', *PLoS Computational Biology*, 7(11), pp. 1–9. doi: 10.1371/journal.pcbi.1002255.
Melorose, J., Perroy, R. and Careas, S. (2015) 'World population prospects', *United Nations*, 1(6042), pp. 587–92. doi: 10.1017/CBO9781107415324.004.
Mendiburu, A. O. and Peloquin, S. J. (1977) 'The significance of 2N gametes in potato breeding', *Theoretical and Applied Genetics*, 49(2), pp. 53–61. doi: 10.1007/BF00275164.
Menéndez, C. M., Ritter, E., Schäfer-Pregl, R., Walkemeier, B., Kalde, A., Salamini, F. and Gebhardt, C. (2002) 'Cold sweetening in diploid potato: Mapping quantitative trait loci and candidate genes', *Genetics*, 162(3), pp. 1423–1434.
Meuwissen, T. and Goddard, M. (2010) 'Accurate prediction of genetic values for complex traits by whole-genome resequencing', *Genetics*, 185(2), pp. 623–631. doi: 10.1534/genetics.110.116590.
Meuwissen, T. H. E. (2009) 'Accuracy of breeding values of "unrelated" individuals predicted by dense SNP genotyping.', *Genetics, selection, evolution : GSE*, 41, p. 35. doi: 10.1186/1297-9686-41-35.
Meuwissen, T. H. E., Hayes, B. J. and Goddard, M. E. (2001) 'Prediction of total genetic value using genome-wide dense marker maps', *Genetics*, 157(4), pp. 1819–1829. doi: 11290733.
Milbourne, D., Meyer, R. C., Collins, a J., Ramsay, L. D., Gebhardt, C. and Waugh, R. (1998) 'Isolation, characterisation and mapping of simple sequence repeat loci in potato.', *Molecular & general genetics : MGG*, 259(3), pp. 233–245. doi: 10.1007/s004380050809.
Milczarek, D., Flis, B. and Przetakiewicz, A. (2011) 'Suitability of Molecular Markers for Selection of Potatoes Resistant to Globodera spp', *American Journal of Potato Research*, 88(3), pp. 245–255. doi: 10.1007/s12230-011-9189-0.
Muir, W. M. (2007) 'Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters', *Journal of Animal Breeding and Genetics*, 124(6), pp. 342–355. doi: 10.1111/j.1439-0388.2007.00700.x.
Müller-Röber, B. T., Koßmann, J., Hannah, L. C., Willmitzer, L. and Sonnewald, U. (1990) 'One of two different ADP-glucose pyrophosphorylase genes from potato responds strongly to elevated levels of sucrose', *MGG Molecular & General Genetics*, 224(1), pp. 136–146. doi: 10.1007/BF00259460.
Naess, S. K., Bradeen, J. M., Wielgus, S. M., Haberlach, G. T., McGrath, J. M. and Helgeson, J. P. (2000) 'Resistance to late blight in Solanum bulbocastanum is mapped to chromosome 8', *TAG Theoretical and Applied Genetics*, 101(5–6), pp. 697–704. doi: 10.1007/s001220051533.
Naik, S. N., Goud, V. V., Rout, P. K. and Dalai, A. K. (2010) 'Production of first and second generation biofuels: A comprehensive review', *Renewable and Sustainable Energy Reviews*, 14(2), pp. 578–597. doi: 10.1016/j.rser.2009.10.003.
Oberhagemann, P., Chatot-Balandras, C., Schäfer-Pregl, R., Wegener, D., Palomino, C., Salamini, F., Bonnel, E. and Gebhardt, C. (1999) 'A genetic analysis of quantitative resistance to late blight in potato: Towards marker-assisted selection', *Molecular Breeding*, 5(5), pp. 399–415. doi: 10.1023/A:1009623212180.
Ortega, F. and Lopez-Vizcon, C. (2012) 'Application of Molecular Marker-Assisted Selection (MAS) for Disease Resistance in a Practical Potato Breeding Programme', *Potato Research*, 55(1), pp. 1–13. doi: 10.1007/s11540-011-9202-5.
Ottoman, R. J., Hane, D. C., Brown, C. R., Yilma, S., James, S. R., Mosley, A. R., Crosslin, J. M. and Vales, M. I. (2009) 'Validation and implementation of marker-assisted selection (MAS) for PVY resistance (Ryadg gene) in a tetraploid potato breeding program', *American Journal of Potato Research*, 86(4), pp. 304–314. doi: 10.1007/s12230-009-9084-0.
Pereira, A. da S., Tai, G. C. C., Yada, R. Y., Coffin, R. H. and Souza-Machado, V. (1994) 'Potential for improvement by

selection for reducing sugar content after cold storage for three potato populations', *Theoretical and Applied Genetics*, 88(6–7), pp. 678–684. doi: 10.1007/BF01253970.
Plaisted, R. L., Bonierbale, M. W., G.C., Y., Pineda, O., Tingey, W. M., Berg, J. van den, Ewing, E. E. and Brodie, B. B. (1994) 'Potato improvement by traditional breeding and opportunities for new technologies', in Belknap, W. R., Vayda, M. E., and Park, W. D. (eds) *The molecular and cellular biology of the potato*, pp. 1–22.
Plaisted, R. L., Sanford, L., Federer, W. T., Kehr, A. E. and Peterson, L. C. (1962) 'Specific and general combining ability for yield in potatoes', *Americal Potato Journal*, 39(397), pp. 185–197.
Poland, J., Endelman, J., Dawson, J., Rutkoski, J., Wu, S. Y., Manes, Y., Dreisigacker, S., Crossa, J., Sanchez-Villeda, H., Sorrells, M. and Jannink, J. L. (2012) 'Genomic Selection in Wheat Breeding using Genotyping-by-Sequencing', *Plant Genome*, 5(3), pp. 103–113. doi: Doi 10.3835/Plantgenome2012.06.0006.
Poland, J. a and Rife, T. W. (2012) 'Genotyping-by-Sequencing for Plant Breeding and Genetics', *The Plant Genome Journal*, 5(3), pp. 92–102. doi: 10.3835/plantgenome2012.05.0005.
Potato Genome Sequencing Consortium, Xu, X., Pan, S., Cheng, S., Zhang, B., Mu, D., Ni, P., Zhang, G., Yang, S., Li, R., Wang, J., Orjeda, G., Guzman, F., Torres, M., Lozano, R., Ponce, O., Martinez, D., De la Cruz, G., Chakrabarti, S. K., Patil, V. U., Skryabin, K. G., Kuznetsov, B. B., Ravin, N. V, Kolganova, T. V, Beletsky, A. V, Mardanov, A. V, Di Genova, A., Bolser, D. M., Martin, D. M. A., Li, G., Yang, Y., Kuang, H., Hu, Q., Xiong, X., Bishop, G. J., Sagredo, B., Mejía, N., Zagorski, W., Gromadka, R., Gawor, J., Szczesny, P., Huang, S., Zhang, Z., Liang, C., He, J., Li, Y., He, Y., Xu, J., Zhang, Y., Xie, B., Du, Y., Qu, D., Bonierbale, M., Ghislain, M., Herrera, M. del R., Giuliano, G., Pietrella, M., Perrotta, G., Facella, P., O'Brien, K., Feingold, S. E., Barreiro, L. E., Massa, G. A., Diambra, L., Whitty, B. R., Vaillancourt, B., Lin, H., Massa, A. N., Geoffroy, M., Lundback, S., DellaPenna, D., Buell, C. R., Sharma, S. K., Marshall, D. F., Waugh, R., Bryan, G. J., Destefanis, M., Nagy, I., Milbourne, D., Thomson, S. J., Fiers, M., Jacobs, J. M. E., Nielsen, K. L., Sønderkær, M., Iovene, M., Torres, G. A., Jiang, J., Veilleux, R. E., Bachem, C. W. B., de Boer, J., Borm, T., Kloosterman, B., van Eck, H., Datema, E., Hekkert, B. te L., Goverse, A., van Ham, R. C. H. J. and Visser, R. G. F. (2011) 'Genome sequence and analysis of the tuber crop potato.', *Nature*, 475(7355), pp. 189–95. doi: 10.1038/nature10158.
R Core Team (2015) 'R: A Language and Environment for Statistical Computing'. Vienna, Austria: R Foundation for Statistical Computing.
Ramakrishnan, A. P., Ritland, C. E., Blas Sevillano, R. H. and Riseman, A. (2015) 'Review of Potato Molecular Markers to Enhance Trait Selection', *American Journal of Potato Research*, 92(4), pp. 455–472. doi: 10.1007/s12230-015-9455-7.
Reid, A., Hof, L., Felix, G., Rücker, B., Tams, S., Milczynska, E., Esselink, D., Uenk, G., Vosman, B. and Weitz, A. (2011) 'Construction of an integrated microsatellite and key morphological characteristic database of potato varieties on the EU common catalogue', *Euphytica*, 182(2), pp. 239–249. doi: 10.1007/s10681-011-0462-6.
Resende, M. F. R., Munoz, P., Resende, M. D. V., Garrick, D. J., Fernando, R. L., Davis, J. M., Jokela, E. J., Martin, T. a., Peter, G. F. and Kirst, M. (2012) 'Accuracy of Genomic Selection Methods in a Standard Data Set of Loblolly Pine (Pinus taeda L.)', *Genetics*, 190(4), pp. 1503–1510. doi: 10.1534/genetics.111.137026.
Riedelsheimer, C., Czedik-Eysenberg, A., Grieder, C., Lisec, J., Technow, F., Sulpice, R., Altmann, T., Stitt, M., Willmitzer, L. and Melchinger, A. E. (2012) 'Genomic and metabolic prediction of complex heterotic traits in hybrid maize', *Nature Genetics*. Nature Publishing Group, 44(2), pp. 217–220. doi: 10.1038/ng.1033.
Rizza, M. D., Vilar, F. L., Torres, D. G. and Maeso, D. (2006) 'Detection of PVY Extreme Resistance Genes in Potato Germplasm from the Urnguayan Breeding Program', *American Journal of Potato Research*, 83(4), pp. 297–304.
Schreiber, L., Nader-Nieto, A. C., Schönhals, E. M., Walkemeier, B. and Gebhardt, C. (2014) 'SNPs in genes functional in starch-sugar interconversion associate with natural variation of tuber starch and sugar content of potato (Solanum tuberosum L.).', *G3 (Bethesda, Md.)*, 4(10), pp. 1797–811. doi: 10.1534/g3.114.012377.
Schultz, L., Cogan, N. O. I., Mclean, K., Dale, M. F. B., Bryan, G. J., Forster, J. W. and Slater, A. T. (2012) 'Evaluation and implementation of a potential diagnostic molecular marker for H1-conferred potato cyst nematode resistance in potato (Solanum tuberosum L.)', *Plant Breeding*, 131(2), pp. 315–321. doi: 10.1111/j.1439-0523.2012.01949.x.
Schäfer-Pregl, R., Ritter, E., Concilio, L., Hesselbach, J., Lovatti, L., Walkemeier, B., Thelen, H., Salamini, F. and Gebhardt, C. (1998) 'Analysis of quantitative trait loci (QTLs) and quantitative trait alleles (QTAs) for potato tuber yield and starch content', *Theoretical and Applied Genetics*, 97(5–6), pp. 834–846. doi: 10.1007/s001220050963.
Schönhals, E. M., Ortega, F., Barandalla, L., Aragones, A., Ruiz de Galarreta, J. I., Liao, J. C., Sanetomo, R., Walkemeier, B., Tacke, E., Ritter, E. and Gebhardt, C. (2016) 'Identification and reproducibility of diagnostic DNA markers for tuber starch and yield optimization in a novel association mapping population of potato (Solanum tuberosum L.)', *Theoretical and Applied Genetics*. Springer Berlin Heidelberg, 129(4), pp. 1–19. doi: 10.1007/s00122-016-2665-7.
Shallenberger, R., Smith, O. and Treadway, R. (1959) 'Food Color Changes - Role of the Sugars in the Browning Reaction in Potato Chips', *Journal of Agricultural and Food Chemistry*, 7(4), pp. 274–277. doi: 10.1021/jf60098a010.
Slater, A. T., Cogan, N. O. I., Forster, J. W., Hayes, B. J. and Daetwyler, H. D. (2016) 'Improving Genetic Gain with Genomic Selection in Autotetraploid Potato', *The Plant Genome*, 0(0), p. 0. doi: 10.3835/plantgenome2016.02.0021.
Slater, A. T., Cogan, N. O. I., Hayes, B. J., Schultz, L., Dale, M. F. B., Bryan, G. J. and Forster, J. W. (2014) 'Improving breeding efficiency in potato using molecular and quantitative genetics', *Theoretical and Applied Genetics*, 127(11), pp. 2279–2292. doi: 10.1007/s00122-014-2386-8.

Slater, A. T., Wilson, G. M., Cogan, N. O. I., Forster, J. W. and Hayes, B. J. (2014) 'Improving the analysis of low heritability complex traits for enhanced genetic gain in potato', *Theoretical and Applied Genetics*, 127(4), pp. 809–820. doi: 10.1007/s00122-013-2258-7.

Śliwka, J., Jakuczun, H., Chmielarz, M., Hara-Skrzypiec, A., Tomczyńska, I., Kilian, A. and Zimnoch-Guzowska, E. (2012) 'A resistance gene against potato late blight originating from Solanum × michoacanum maps to potato chromosome VII', *Theoretical and Applied Genetics*, 124(2), p. 397406. doi: 10.1007/s00122-011-1715-4.

Song, J., Bradeen, J. M., Naess, S. K., Raasch, J. a, Wielgus, S. M., Haberlach, G. T., Liu, J., Kuang, H., Austin-Phillips, S., Buell, C. R., Helgeson, J. P. and Jiang, J. (2003) 'Gene RB cloned from Solanum bulbocastanum confers broad spectrum resistance to potato late blight.', *Proceedings of the National Academy of Sciences of the United States of America*, 100(16), pp. 9128–9133. doi: 10.1073/pnas.1533501100.

Sowokinos, J. R. (2001) 'Biochemical and molecular control of cold-induced sweetening in potatoes', *American Journal of Potato Research*, 78, pp. 221–236. doi: 10.1007/BF02883548.

Townsend, L. R. and Hope, G. W. (1960) 'Factors influencing the colour of potato chips', *Canadian Journal of Plant Science*, 40, pp. 58–64.

Uitdewilligen, J. G. A. M. L., Wolters, A. M. A., D'hoop, B. B., Borm, T. J. A., Visser, R. G. F. and van Eck, H. J. (2013) 'A Next-Generation Sequencing Method for Genotyping-by-Sequencing of Highly Heterozygous Autotetraploid Potato', *PLoS ONE*, 8(5), pp. 10–14. doi: 10.1371/journal.pone.0062355.

Umaerus, V. and Umaerus, M. (1994) 'Inheritance of resistance to late blight', in Bradshaw, J. E. and Mackay, G. R. (eds) *Potato genetics1*. CAP International, pp. 365–401.

Urbany, C., Stich, B., Schmidt, L., Simon, L., Berding, H., Junghans, H., Niehoff, K.-H., Braun, A., Tacke, E., Hofferbert, H.-R., Lübeck, J., Strahwald, J. and Gebhardt, C. (2011) 'Association genetics in Solanum tuberosum provides new insights into potato tuber bruising and enzymatic tissue discoloration.', *BMC genomics*, 12(1), p. 7. doi: 10.1186/1471-2164-12-7.

Valin, H., Sands, R. D., van der Mensbrugghe, D., Nelson, G. C., Ahammad, H., Blanc, E., Bodirsky, B., Fujimori, S., Hasegawa, T., Havlik, P., Heyhoe, E., Kyle, P., Mason-D'Croz, D., Paltsev, S., Rolinski, S., Tabeau, A., van Meijl, H., von Lampe, M. and Willenbockel, D. (2013) 'The future of food demand: understanding differences in global economic models', *Agricultural Economics*, 45, p. n/a–n/a. doi: 10.1111/agec.12089.

VanRaden, P. M. (2008) 'Efficient methods to compute genomic predictions.', *Journal of dairy science*. Elsevier, 91(11), pp. 4414–23. doi: 10.3168/jds.2007-0980.

VanRaden, P. M., Van Tassell, C. P., Wiggans, G. R., Sonstegard, T. S., Schnabel, R. D., Taylor, J. F. and Schenkel, F. S. (2009) 'Invited review: reliability of genomic predictions for North American Holstein bulls.', *Journal of dairy science*. Elsevier, 92(1), pp. 16–24. doi: 10.3168/jds.2008-1514.

Wellmann, R., Preuß, S., Tholen, E., Heinkel, J., Wimmers, K. and Bennewitz, J. (2013) 'Genomic selection using low density marker panels with application to a sire line in pigs.', *Genetics, selection, evolution : GSE*, 45, p. 28. doi: 10.1186/1297-9686-45-28.

Wetterstrand, K. (2017) *DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP)*. Available at: www.genome.gov/sequencingcostsdata (Accessed: 8 March 2017).

Wiggans, G. R., VanRaden, P. M. and Cooper, T. A. (2011) 'The genomic evaluation system in the United States: Past, present, future', *Journal of Dairy Science*. Elsevier, 94(6), pp. 3202–3211. doi: 10.3168/jds.2010-3866.

Wolc, A., Stricker, C., Arango, J., Settar, P., Fulton, J. E., O'Sullivan, N. P., Preisinger, R., Habier, D., Fernando, R., Garrick, D. J., Lamont, S. J. and Dekkers, J. C. (2011) 'Breeding value prediction for production traits in layer chickens using pedigree or genomic relationships in a reduced animal model', *Genetics Selection Evolution*, 43(1), p. 5. doi: 10.1186/1297-9686-43-5.

Wolc, A., Zhao, H. H., Arango, J., Settar, P., Fulton, J. E., O'Sullivan, N. P., Preisinger, R., Stricker, C., Habier, D., Fernando, R. L., Garrick, D. J., Lamont, S. J. and Dekkers, J. C. M. (2015) 'Response and inbreeding from a genomic selection experiment in layer chickens.', *Genetics, selection, evolution : GSE*. Genetics Selection Evolution, 47(1), p. 59. doi: 10.1186/s12711-015-0133-5.

PLANNED PUBLICATIONS

Planned publications

The work presented in this thesis, excluding results included in Paper 1 and Paper 2, is planned to be submitted to peer review journals. Below is a list of working titles for planned publications along with short descriptions.

Comparison of bulk segregant analysis and genomic selection for prediction and association of dry matter content in tetraploid potato

Sverrisdóttir, E., Johnsen, H. Ø., Nielsen, K.L.

Molecular breeding has focused on MAS, where DNA markers selected to have highest predictive power for relevant traits are used for selection of potential breeding candidates. In contrast, GS is a genome-wide marker approach, where all markers are employed without selection. It is presently unknown which approach is the more powerful, and a comparison is thus of interest. Results in this thesis showed that while GBS data were significantly less powerful for identifying significant markers than BSA, prediction models performed with specifically selected markers did not result in any gain in prediction accuracy compared to when using randomly selected markers, suggesting that a genome-wide approach is just as powerful as MAS. GWAS results also presented in the thesis will possibly be included in the paper.

Genomic prediction of tuber yield in tetraploid potato

Sverrisdóttir, E., Nielsen, K.L.

While potato tuber yield is of significant importance for improved breeding gains, predictions of yield are challenging due to low heritability, strong environmental influence, and non-additive genetic effects. In this thesis, genomic predictions of yield were significantly poorer than predictions of dry matter and chipping quality. Predictions could possibly be improved by use of nonparametric prediction models, and results from this will be included in the paper.

Genomic prediction of late blight resistance in tetraploid potato

Sverrisdóttir, E., Nielsen, K.L.

Late blight is a devastating disease for potato, and resistance against late blight is thus of great significance for a sustainable potato production, both economically and environmentally. In an attempt to separate dominant and quantitative resistance, two models were constructed for this thesis, and although a difference was seen in prediction accuracies, more studies are needed to elucidate the differences, and in particular, to construct better and more robust prediction models for mainly the quantitative late blight resistance.

Optimising genomic prediction in tetraploid potato

Sverrisdóttir, E., Nielsen, K.L.

As was seen from the results presented, a number of different factors influence the accuracy of genomic prediction models, such as trait heritability, size of training population, and genetic relationship between individuals. Of particular interest is the number of markers required for robust and unbiased predictions with GS. Results suggested that 10,000 markers were sufficient to obtain good prediction accuracies for dry matter and chipping quality, however, it would be interesting to see if this number could be reduced when removing all missing data, or if perhaps the predictions could be improved if all individuals had data for the same markers. Furthermore, as seen in Paper 2, predictions conducted with markers selected randomly performed similarly compared to when markers were selected based on amount of missing data. This was also indicated by predictions performed with both BSA and GWAS selected markers. More analyses are needed to gain insight to the influence of marker number. All of this will be collected in a paper describing more general aspects surrounding genomic selection and how these can be utilised for optimal genomic prediction models in tetraploid potato.

Breeding for space and resource efficient crops is more important than ever to feed the world's increasing population. Potatoes produce twice the amount of calories per area compared to cereals and are thus of central importance for global food security. While traditional potato breeding is costly and time-consuming, molecular breeding techniques such as genomic selection have the potential to speed up the breeding process significantly. Genomic selection uses genome-wide molecular markers to predict performance of individuals, and selection of breeding candidates can be performed without direct phenotyping. Genomic selection has been implemented in breeding programs for dairy cattle, and a number of studies report promising results in both animal and plant species. The overall objective of this thesis was to evaluate the potential of and initiate a genomic selection breeding programme in tetraploid potato. Implementation of genomic selection in potato breeding programmes has the potential to accelerate the genetic gain significantly, giving potato a prominent role in ensuring food security in the future.

Elsa Sverrisdóttir
Phone: +45 50 55 30 92
E-mail: esv@bio.aau.dk | elsasverris@gmail.com
LinkedIn: dk.linkedin.com/in/elsasverris/

ISSN (online): 2446-1636
ISBN (online): 978-87-7112-924-3

AALBORG UNIVERSITY
DENMARK

AALBORG UNIVERSITY PRESS